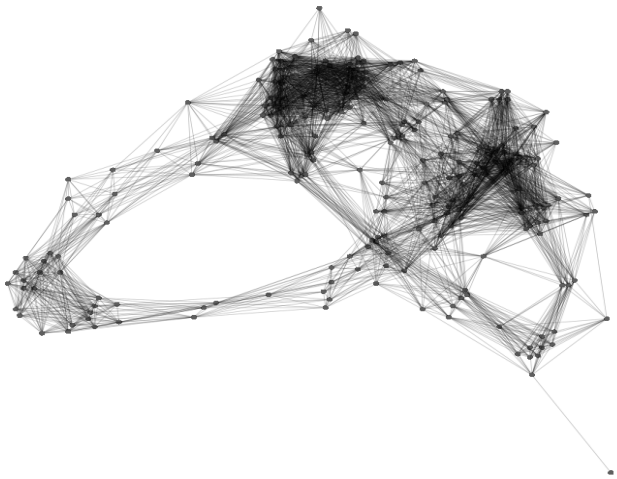
Facultad de Ciencias

Campus de Puerto Real

Análisis de secuencias repetidas de ADN en el lenguado *Solea senegalensis* mediante tratamiento de datos procedentes de NGS (Next Generation Sequencing)

Grado en Biotecnología

Curso académico 2017/2018



AUTOR: ÁLVARO BENÍTEZ MATEO

TUTOR: ISMAEL CROSS PACHECO

Gracias a mi tutor, Ismael,

por ser mi guía y fuente de inspiración en todo este proyecto.

A mi familia y amigos,

por apoyarme siempre que ha sido necesario.

ÍNDICE

RESUMEN

Los elementos repetitivos conforman un gran porcentaje del genoma de los seres vivos. Además, suponen un enorme impacto en la regulación génica, la inactivación y eliminación de genes, en la estructura cromosómica y en la evolución de los genomas. Por otra parte, el rápido desarrollo de las tecnologías de secuenciación masiva de genomas permite obtener una gran cantidad de información que ha de ser tratada con herramientas bioinformáticas para poder ser anotada y organizada. Por todo ello, se han analizado cuatro paquetes de datos procedentes de la secuenciación del transcriptoma de *Solea senegalensis* mediante la plataforma *Galaxy-RepeatExplorer*, siendo esta una especie de pez de gran importancia económica y que está siendo estudiada a fin de solucionar los problemas que rodean su cría en cautividad.

Los análisis de los datos han permitido obtener más de 2.500 *clusters* indicativos de posibles elementos repetitivos, de los cuales se han analizado los 32 más significativos. Así, se han encontrado varios elementos que pueden constituir satélites y retrotransposones que no se hallaban en las bases de datos consultadas, mostrándose este como un sistema muy eficiente de identificación y caracterización de posibles elementos repetitivos *de novo* en especies no modelo. Además, una vez establecido el proceso, se ha generado un diagrama de trabajo o *workflow* que recoge los diferentes pasos llevados a cabo con las herramientas disponibles en la plataforma *Galaxy-RepeatExplorer* para llegar a obtener los resultados mostrados.

ABSTRACT

Repetitive elements constitute a large percentage of the genome of the living thing genomes. In addition, they have a huge impact on gene regulation, gene inactivation and deletion, chromosome structure and genome evolution. On the other hand, the rapid development of next genome sequencing technologies makes it possible to obtain a large amount of information that needs to be treated with bioinformatics tools in order to be recorded and organized. Therefore, four data packages from the sequencing of the *Solea senegalensis* transcriptome using the *Galaxy-RepeatExplorer* platform have been analysed. This is a species of fish of great economic importance and is being studied in order to solve the problems related with its captive breeding.

Data analysis has allowed to compilate over 2500 indicative clusters of possible repetitive elements, the 32 most important of these have been analyzed. Thus, several elements have been found that can constitute satellites and retrotransposons that were not found in the databases consulted, showing this as a very efficient system of identification and characterization of possible repetitive elements de novo in non-model species in a fast way. In addition, once the process has been established, a working diagram or workflow has been generated that gathers the different steps carried out with the tools available in the *Galaxy-RepeatExplorer* platform to obtain the results shown.

1. INTRODUCCIÓN

1.1. Descripción y taxonomía de la especie de estudio

Solea senegalensis o lenguado común (Fig.1) es una especie de la familia Soleidae, del orden de los Pleuronectiformes, clase Actinopterygii, filo Cordata y reino Animalia. Dentro de esta familia, se encuentra la especie *Solea solea*, también conocida como lenguado común, con la cual comparte muchas características biológicas y morfológicas (Colen *et al.*, 2018). La forma de diferenciarlas visualmente requiere cierta experiencia, basándose esta identificación en la observación de la membrana interradial de la aleta pectoral del lado ocular, la cual presenta una coloración negra en el caso de *S. senegalensis*, frente a una mancha negra compacta en *S. solea* (Colen *et al.*, 2014; Imsland, 2010).

Figura 1. Fotografía de lenguados creciendo en tanque (Colen et al., 2014).

1.2. Impacto económico de la especie *S. senegalensis*

La importancia comercial de esta especie está ligada a la especie de lenguado común *S. solea*. A pesar de poseer características físicas similares, su consumo humano está polarizado, estando mejor valorada la especie *S. solea* en países del norte y del este de Europa, mientras que, en el sur del continente, sobre todo en España y Portugal, tiene mayor importancia la especie *S. senegalensis* (Imsland, 2010). Siendo así, se ha pasado de una producción global de 347 y 571 toneladas en los años 2010 y 2013 (Colen *et al.*, 2014), a producirse 1.500 toneladas en 2016 (APROMAR, 2017), con España y Francia como países líderes en la producción de lenguado a nivel mundial. No obstante, aunque en España y Portugal hay un gran interés por la especie *S. senegalensis*, su producción en la acuicultura aún sigue teniendo diversos problemas que no se han terminado de solucionar, por lo que está sujeto a estudios en estos países (Dinis, 1999). Entre estos problemas destaca la infertilidad presente en los individuos crecidos en cautividad. Esta infertilidad, presente tanto en machos como hembras, se ha tratado de solucionar con tratamientos hormonales sin obtener resultados satisfactorios (Cabrita *et al.*,

2006). En los individuos machos de la F1 se observa cómo se produce una correcta espermatogénesis y presentan buenos niveles de andrógenos, sin embargo, el problema que provoca la infertilidad es una reducida producción de esperma frente a los individuos salvajes capturados (Cabrita *et al.*, 2006; Guzmán *et al.*, 2011). Se ha tratado de solucionar este problema mediante el tratamiento con agonistas liberadores de hormona gonadotropina (GnRHa), mostrándose como un tratamiento ineficaz en la mejora de la infertilidad. Del mismo modo, se han realizado estudios a fin de comparar el tratamiento con GnRHa con el tratamiento mediante gonadotropina coriónica humana (hCG) y pese a observarse una mejora y un aumento en la fertilidad, no se considera un tratamiento capaz de eliminar el problema de la infertilidad en individuos de la F1 (Guzmán *et al.*, 2011).

La producción acuícola de *S. senegalensis* se ve condicionada por su ciclo de vida, caracterizado por la metamorfosis que comienza en torno al día 11 después de la eclosión y termina ocho días después, el día 19 después de la eclosión (Dinis, 1999) y por el conocimiento de los mecanismos de control sexual en la especie.

1.3. Metamorfosis en peces planos

El proceso de metamorfosis está presente en los peces planos, los cuales sufren un gran cambio al desarrollarse: su cuerpo gira 90º y uno de los ojos migra de un lado a otro, quedando ambos ojos en el lado derecho tras la migración en el caso del lenguado senegalés. Dicha metamorfosis adapta al pez para su vida bentónica, conllevando grandes cambios en su fisiología, hábitos alimentarios, pigmentación, comportamiento y morfología (Fernández *et al.*, 2017; Laudet *et al.*, 2011). Las condiciones que influyen en la metamorfosis, ya sea el tamaño, edad, o condiciones ambientales como la temperatura o el alimento disponible, constituyen una importante fuente de información para el desarrollo y la mejora de la cría en cautividad de estas especies, afectando a la productividad y la eficacia de dicha cría (Fernández-Díaz *et al.*, 2001). Actualmente se están enfocando estos estudios hacia la influencia en el desarrollo metamórfico de la presencia de determinados nutrientes como la vitamina A y la melatonina (Fernández *et al.*, 2017; Lan-Chow-Wing *et al*., 2014). Se ha estudiado cómo en los pleuronectiformes la metamorfosis parece en gran medida estar controlada por las hormonas tiroideas (Laudet *et al.*, 2011). Así mismo, existen genes como los *HSP90*, los cuales codifican para chaperonas de gran importancia en la regulación de los ciclos celulares, que se encuentran regulados por las hormonas tiroideas, remarcando la importancia de estas en el desarrollo del individuo (Manchado *et al.*, 2008).

1.4. Mecanismos de determinación sexual

Especialmente importante para la cría en cautividad y el crecimiento en piscifactorías de *S. senegalensis* son los mecanismos de diferenciación sexual. Se conocen sistemas de diferenciación sexual basados en factores externos, en causas ambientales como la temperatura, tamaño, edad o incluso presencia de parásitos (Baroiller *et al.*, 2009). Por otra parte, están los mecanismos genéticos multifactoriales, los cuales pueden determinar el sexo del individuo en base a la presencia de diferentes factores en el genoma. De esta manera, un mayor número de factores inductores de un sexo será lo que lo determine. Por el contrario, mecanismos genéticos basados en un único factor determinarán el sexo según los genes presentes en un único locus (Heule *et al*., 2014).

De entre todos los mecanismos para la determinación del sexo, los más ampliamente observados en la naturaleza y los mejor conocidos son los sistemas XX-XY y WZ-ZZ, donde los individuos homogaméticos son hembras y machos, respectivamente (Chalopin *et al.*, 2015). Particularmente, *S. senegalensis* posee 21 pares de cromosomas, y carece de cromosomas sexuales heteromórficos que puedan ser identificados por su morfología, no obstante, se estudia la posibilidad de que se pueda diferenciar el sistema XX-XY (Molina-Luzón *et al.*, 2015). Recientemente se ha planteado la posibilidad de que dentro del cariotipo de *S. senegalensis*, el cromosoma metacéntrico de mayor tamaño suponga un proto-cromosoma sexual. Esto se deduce de la presencia de los genes *dmrt1-dmrt2-dmrt3* en dicho cromosoma, sabiéndose que estos genes están relacionados con el proceso de determinación del sexo, encontrándose por ejemplo en el cromosoma Z de la especie *Cynoglossus semilaevis* (Portela-Bens *et al.*, 2017).

Es en estos sistemas heterogaméticos en los que tiene lugar un proceso de cambio degenerativo en el cromosoma determinante del sexo, a fin de evitar que se produzca una recombinación meiótica con el otro cromosoma sexual (Charlesworth *et al.*, 2005). Durante este proceso suceden un gran número de mutaciones, que provocan deleciones en genes funcionales, y se incrementa el número de elementos transponibles (TEs) y de ADN repetitivo.

Toda esta evolución observada en los cromosomas sexuales está controlada tanto por la presencia y acumulación de elementos transponibles, que funcionan alterando el contenido génico y la estructura de los cromosomas Y y W, como la regulación génica, silenciando y condensando el material genético. No obstante, aunque se conoce la alta concentración de elementos transponibles en los cromosomas sexuales, no se conoce aún si intervienen de manera activa o pasiva en la evolución de dichos cromosomas. Actualmente se está también estudiando el funcionamiento y la forma que tienen de evolucionar los cromosomas sexuales, a fin de descubrir si una región incipiente en la determinación del sexo hace que haya una alta concentración de elementos transponibles o por el contrario es la alta concentración de

elementos transponibles la que hace que una región sea determinante del sexo (Chalopin *et al.*, 2015).

1.5. Ciencias "*-ómicas*" e impacto de las tecnologías NGS

Es muy importante también el papel de las denominadas "-ómicas" en muchos de los avances de los últimos años. La llamada genómica hace referencia a la ciencia que estudia a gran escala los genomas. Además, han surgido multitud de ciencias derivadas, tales como la transcriptómica, el estudio de todos los transcritos de ARN en un organismo en unas condiciones determinadas; la proteómica, la ciencia que estudia las proteínas del organismo, etc. Estas ciencias se han podido desarrollar gracias a los avances en la secuenciación masiva de ADN mediante las tecnologías NGS (Next Generation Sequencing) y la bioinformática (Xu *et al.* 2006).

Las NGS han abierto todo un abanico de posibilidades en investigación, permitiendo obtener enormes volúmenes de datos a muy bajo coste. Algunas de las NGS mejor implantadas comercialmente son las de segunda generación como Illumina de Solexa y 454 de Roche, con un gran número de especies y genomas analizados en los últimos años. En la actualidad, las NGS de tercera generación están implantándose y desplazando a algunas de segunda generación tomando gran fuerza en proyectos de secuenciación *de novo* de genomas más complejos.

Estas tecnologías tienen en común un primer paso de rotura del ADN, generando fragmentos pequeños del material genético. Estos fragmentos son amplificados generalmente mediante una PCR (Reacción en Cadena de la Polimerasa) en emulsión (Dressman *et al.*, 2003) o con una amplificación en fase sólida (Fedurco *et al.*, 2006). Una vez que se ha amplificado el ADN y se tienen millones de copias de cada fragmento, se procede a determinar las secuencias de estos. Para ello, diferentes tecnologías usan diferentes métodos. En Roche 454 se realiza la pirosecuenciación que consiste en la adición de nucleótidos en un orden determinado y en bucle, y conforme se unen a la secuencia de ADN fluoresce y es detectado por una cámara o detector. En el caso de la tecnología Illumina, se adicionan nucleótidos marcados por fluorescencia con terminaciones reversibles, de manera que en cada ciclo se añaden los cuatro nucleótidos (A, T, G, C). Cada nucleótido unido libera una señal de diferente color que es detectada y anotada en la secuencia. Esto genera una gran cantidad de datos conocidos como *reads*, que son las lecturas de los pequeños fragmentos de ADN secuenciados. Estas secuencias de ADN, que pueden variar en tamaño según la tecnología de secuenciación utilizada, son analizadas posteriormente mediante proogramas bioinformáticos. A partir de ellos se puede reconstruir la secuencia inicial de los genomas mediante su ensamblaje, analizar genes transcritos, estudiar polimorfismos o describir secuencias repetidas como los elementos transponibles o los SSRs (*Simple Sequence Repeats*) (Metzker, 2010).

Actualmente a pesar de los grandes avances alcanzados, uno de los mayores cuellos de botella para estas ciencias es la falta de herramientas informáticas que permitan el procesado de datos masivos a investigadores más enfocados al campo de la biología y menos especializados en la informática. Además, la mayor parte del software disponible para el análisis genómico requiere genomas completamente ensamblados, en los cuales a menudo se eliminan las regiones repetitivas, de las que se hablará en el siguiente apartado, que suponen un problema durante la purificación de los *reads* y el posterior ensamblaje, haciendo de esto un impedimento a la hora de estudiar las repeticiones del genoma.

Del mismo modo, otro problema para el análisis de repeticiones es que la mayor parte del *software* disponible detecta estas repeticiones al compararlas con bases de datos de secuencias repetitivas consenso, imposibilitando así la detección de repeticiones nuevas que no se vean identificadas en esas bases de datos (Novak *et al.*, 2013).

1.6. Elementos repetitivos: Concepto y clasificación

Dentro de los genomas eucariotas los elementos repetitivos son un componente mayoritario y de gran importancia que contribuyen al tamaño, diversidad y estructura de dichos genomas (Sotero-Caio *et al.*, 2017). Es prioritario conocer sus características y clasificación básica para comprender el trabajo que se expone.

La primera cuantificación de elementos repetitivos realizada en un organismo vertebrado superior fue realizada tras la secuenciación del genoma humano. En un primer momento, se consideró que el genoma humano estaba compuesto por al menos un 50% de ADN repetitivo, siendo un 90% de estas repeticiones elementos transponibles (TEs), tal como se determinó en el proyecto genoma humano según el International Human Genome Sequencing Consortium (2001). No obstante, según estudios más actualizados, se ha encontrado que en el ser humano hasta un 69% del genoma se encuentra formado por elementos repetitivos (de Koning *et al.*, 2011). Esta gran diferencia puede deberse en gran medida a las técnicas de detección de elementos repetitivos utilizadas. Generalmente, se identifican los elementos repetitivos mediante algoritmos que no buscan elementos repetitivos en sí mismos, sino que tratan de comparar los datos introducidos con bases de datos que contienen secuencias consenso de elementos repetitivos, buscando así componentes similares a los que identifican como elementos repetitivos. Sin embargo, con el estudio descrito hace unos años por *de Koning* (2011) se están desarrollando unos tipos de análisis que funcionan buscando las propias repeticiones, y no similitudes con secuencias consenso. Para ello, se buscan en primer lugar pequeñas secuencias de oligonucleótidos que se encuentren muy repetidas en el genoma, tras esto, se buscan oligonucleótidos relacionados que aparezcan agrupados. Así se forman "*nubes-*

P” (*P-clouds*) que identificarán regiones del genoma cuyo origen sea de carácter repetitivo (de Koning *et al.*, 2011).

Actualmente se está revisando el papel crucial que juegan los elementos transponibles en la evolución de las especies (Serrato-Capuchina y Matute, 2018), además de otras implicaciones como la estrecha relación entre ciertas familias de repeticiones y la capacidad de adaptación al medio que aportan a muchos organismos, entre ellos a los peces, que las poseen (Yuan *et al.*, 2018). Es importante el papel de las repeticiones en la modificación de los genomas, creando nuevos genes, modificando los ya existentes y reordenando la información presente en los cromosomas, así como su utilidad como marcadores de procesos evolutivos, permitiendo el estudio de procesos de selección y mutación (Oliver y Green, 2009).

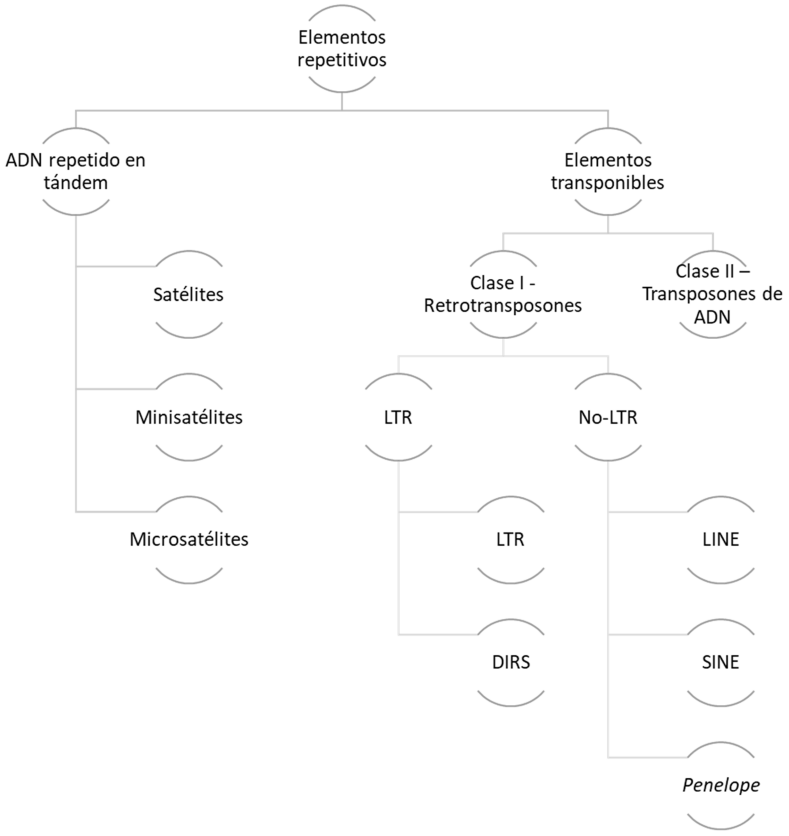


Figura 2. Clasificación de elementos repetitivos.

Clasificación de los elementos repetitivos de los genomas

El ADN repetitivo se divide en ADN repetido en tándem (satélites, minisatélites y microsatélites) y elementos transponibles (Fig. 2) (Jurka *et al.*, 2005; International Human Genome Sequencing Consortium, 2001). Las repeticiones en tándem están siendo muy estudiadas actualmente, dado la gran información que aportan los polimorfismos en su secuencia acerca de relaciones parentales y de pedigrí. Estos estudios se han visto potenciados gracias al avance de las NGS y de herramientas informáticas para la detección de repeticiones

de novo como *RepeatExplorer*. Además, dado que el enorme volumen de copias de una misma repetición en el genoma supone un problema para el tratamiento de datos, se está comenzando a enfocar los estudios de los satélites a lecturas del genoma transcrito (transcriptoma) (Shah *et al.*, 2016).

Por otro lado, se encuentran los elementos transponibles, los cuales son responsables de parte del control epigenético y de la regulación de genes cercanos, además de funcionar como una posible fuente de nuevos genes para los hospedadores (Chalopin *et al.*, 2015). Los TEs pueden clasificarse en dos grandes grupos, diferenciados según su mecanismo para movilizarse en el genoma. En los TEs de clase I se encuentran los retrotranposones, pudiendo diferenciarse los *long terminal repeat* (LTR) y los no-LTR. Esta clase de TE se moviliza mediante el empleo de retrotransciptasas, de manera que el ARN traducido se retrotranscribe a una molécula de ADN de cadena doble, que se integrará posteriormente en el genoma. La forma de integrarse es lo que hace que se diferencien LTR y no-LTR (Sotero-Caio *et al.*, 2017). Además, entre los no-LTR se han de diferenciar diferentes subclases, si son autónomos se denominan LINEs y si no son autónomos dependerán de los anteriores y se denominarán SINEs, añadiendo además una tercera clase denominada *Penelope*. En cuanto a los TEs de clase II, son aquellos que se movilizan sin una transcripción inversa. Generalmente, su presencia es minoritaria en comparación a los TEs de clase I (Sotero-Caio *et al.*, 2017).

1.7. Genómica de *S. senegalensis*

A fin de tratar los problemas relacionados con la producción acuícola, se necesita del mayor número de datos posibles acerca de la especie a nivel genómico, así como analizar y anotar dicha información. En este sentido, en *S. senegalensis* se han identificado marcadores moleculares y citogenéticos (familias multigénicas y secuencias teloméricas) y se han desarrollado genotecas BAC y realizado mapas genéticos integrados (Portela-Bens *et al.*, 2017; García-Cegarra *et al.*, 2013), además de secuenciación *de novo*, caracterización y anotación funcional de su transcriptoma (Benzekri *et al.*, 2014). Actualmente, si bien se sabe que *S. senegalensis* posee 21 parejas de cromosomas (Hardie y Hebert, 2004), su genoma nuclear no se ha secuenciado y únicamente se encuentran mapeados 129 microsatélites (Robledo *et al.*, 2017). Además, está disponible un mapa físico de su genoma (Molina-Luzón *et al.*, 2015).

1.8.Análisis de secuencias repetidas basados en grafos

Para solventar el problema en la infraestimación de elementos repetidos en los genomas, se han desarrollado herramientas informáticas sofisticadas que facilitan el análisis y agrupación de repeticiones de una manera mucho más precisa. Una de ellas está basada en la agrupación de lecturas de secuencias (*reads*) procedentes de NGS mediante el uso de grafos (Novak *et al.*,

2013). Un grafo (del griego *grafos*: dibujo, imagen) es un conjunto de nodos, que en el caso que aquí ocupa representarían cada una de las lecturas, unidos por enlaces o aristas, que permiten representar relaciones binarias entre ellos. En los grafos, la longitud de las aristas representa la similitud entre los *reads*. En un grafo los *reads* se conectarán según su similitud. Aquellos que solapen en mayor medida formarán más conexiones que darán lugar a *contigs* (secuencia resultante de la unión de lecturas). Estos contigs a su vez se pueden agrupar en *cluster*s (Fig.3) (Novak *et al.* 2010).

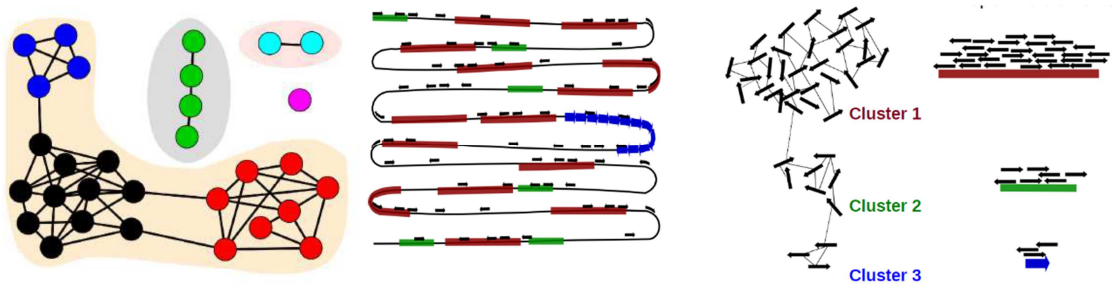


Figura 3. Ejemplo de cómo se organizan los reads en un grafo. Izda: los nodos que comparten color forman contigs y las zonas sombreadas conforman los diferentes clusters. Dcha: Representación esquemática de la obtención de la secuencia de tres tipos de repeticiones con RepeatExplorer.

Los grafos con mayor densidad y con mayor número máximo de líneas que convergen a un mismo nodo (grado máximo) están generalmente relacionados con repeticiones en tándem pequeñas (microsatélites principalmente) y el diámetro del grafo es proporcional a la longitud del elemento repetitivo. De esta forma, el análisis de las formas de los grafos da información sobre el tipo de repetición (TEs, SSRs, etc) que se ha encontrado (Fig. 4) (Novak *et al.*, 2010).

Figura 4. Ejemplo de grafos. Izquierda: Ejemplo de retrotransposón LTR, identificado como GmCL2; Derecha: Ejemplo de repetición en tándem, PsCL21 (modificado de Novak et al., 2010.)

En el presente trabajo se van a analizar paquetes de *reads* del proyecto transcriptoma de *S. senegalensis (Benzekri et al., 2014)* en busca de secuencias repetitivas en el ADN transcrito procedente de diferentes etapas del desarrollo larvario y la metamorfosis, así como las secuencias transcritas relacionadas con el sistema inmune y el desarrollo gonadal de *S. senegalensis*.

2. OBJETIVOS

La abundancia y distribución de los diferentes elementos repetitivos en el genoma de los organismos supone una importante fuente de información de elevado interés. El desarrollo de tecnologías de secuenciación masiva (NGS) amplía las fronteras del estudio de estos elementos en organismos de interés como *S. senegalensis*. Por todo esto, los objetivos que se persiguen con la realización de este trabajo de fin de grado se exponen a continuación:

- Analizar paquetes de datos de secuencias del transcriptoma de *S. senegalensis* para la identificación de elementos repetitivos de ADN mediante herramientas bioinformáticas de alto rendimiento.
- Identificar y caracterizar los transposones y satélites obtenidos tras el análisis de las lecturas obtenidas de las bases de datos del transcriptoma de *S. senegalensis*.
- Comparar los resultados obtenidos de diferentes conjuntos de datos procedentes del transcriptoma obtenido de distintas condiciones y órganos de esta especie: sistema inmune, metamorfosis y gónadas.
- Buscar homologías con genes y secuencias ya anotadas en bases de datos disponibles, mediante algoritmos de comparación.
- Desarrollar un esquema de análisis anidado de herramientas bioinformáticas (*Workflow*) que permita sistematizar este tipo de análisis realizado para ampliarlos a otros conjuntos de datos masivos genómicos disponibles.

3. MATERIAL Y MÉTODOS

3.1. Paquetes de datos de secuenciación

Para el desarrollo del trabajo se han usado parte de las lecturas generadas durante el proyecto transcriptoma de *Solea* (*Benzekri et al., 2014)*. Los *reads* del transcriptoma están organizados en diferentes paquetes según las condiciones del organismo o de la parte de la que se tomó el ARN. De esta manera, de los 51 paquetes disponibles para *S. senegalensis*, se han analizado en esta memoria los mostrados en la Tabla 1:

Tabla 1. Tabla resumen de los paquetes de reads *usados para el trabajo (SRA-NCBI Database: https://www.ncbi.nlm.nih.gov/sra ; Benzekri et al., 2014).*

Tecnología	Condiciones	Código SRR	*Reads* totales
Illumina	Larva en metamorfosis (Etapa S0)	SRR1576631	67.844.392
	Lavar en metamorfosis (Etapa S4)	SRR1576697	63.604.222
454	Gónadas	SRR1581174	731.882
		SRR1581175	664.382
	Órganos relacionados con el sistema inmune	SRR1581197	88.538
		SRR1581199	94.252
		SRR1581200	138.400
		SRR1581201	476.820
		SRR1581203	430.091
TOTAL			134.072.979

Estas secuencias (*reads)* se encuentran alojadas en la base de datos SRA del *National Center for Biotechnology Information* (NCBI), con dirección web: https://www.ncbi.nlm.nih.gov/sra. A ellos se puede acceder con cada código identificativo SRR. En total se han analizado 9 paquetes de datos con más de 130 millones de lecturas de ADN de dos plataformas NGS (454 e Illumina) procedentes de cuatro fuentes diferentes: dos etapas de la metamorfosis, gónadas y órganos del sistema inmune de *S. senegalensis.*

3.2. Análisis de calidad de las lecturas

Para el análisis de calidad de las secuencias se ha utilizado el programa *FastQC*. Este programa (disponible en la plataforma *on-line* de *Galaxy-RepeatExplorer*, con dirección web: https://repeatexplorer-elixir.cerit-sc.cz/) analiza el paquete de *reads* seleccionado y devuelve una serie de datos de especial importancia en el filtrado y el tratamiento de las secuencias. La información de salida aporta datos como la cantidad de secuencias de entrada, la calidad de la secuenciación por bases y por secuencias, la distribución de las longitudes de los fragmentos secuenciados, etc. (Novak *et al.*, 2013).

3.3. Pretratamiento y filtrado de las lecturas

A partir de los datos obtenidos de calidad se puede ejecutar un correcto filtrado de las secuencias mediante la función *Preprocessing of fastq reads* del *Galaxy-Repeat Explorer*. Con esta herramienta se podrán aplicar ciertos criterios con los que filtrar las secuencias disponibles y eliminar las partes de menor calidad. En este sentido, cada base de cada secuencia lleva ligado un valor *Phred:*

Phred score = - 10 log (probabilidad de error en la lectura)

Este índice cuantifica la calidad de la secuenciación con la finalidad de hacer una limpieza de las secuencias para mantener siempre un mínimo de calidad. Esto se controla directamente con los parámetros de límite de calidad y la cantidad de *reads* que han de estar por encima del límite. También se controla la calidad de la secuenciación al cortar los extremos de las secuencias según sea necesario, para ello, se usa como guía la información obtenida con el *FastQC*. De esta manera, el límite inferior supone el corte de las primeras *X* bases de cada secuencia. Análogamente, el límite superior indica a partir de qué posición se eliminan las bases de cada lectura, pero, además, se eliminan por completo aquellos *reads* que no lleguen a la longitud marcada por este.

A modo de ejemplo, en el caso del paquete de datos de larvas en estado S0 de metamorfosis, se retiran todos los *reads* que no alcancen un tamaño de 76 nucleótidos, y posteriormente se eliminan las primeras 15 bases, así como aquellas a partir de la 76.

3.4. Búsqueda y análisis de repeticiones: Plataforma *"RepeatExplorer"*

RepeatExplorer es un conjunto de *softwares* que permiten el análisis de repeticiones en el genoma. Esta plataforma trabaja con una serie de servidores en ordenadores remotos que alojan los programas para el tratamiento masivo de datos mediante supercomputación. Dado que su algoritmo se basa en la selección de pequeños fragmentos del genoma cargados en la plataforma por el usuario, supone una herramienta muy útil para el tratamiento de datos procedentes de la secuenciación de nueva generación (NGS). Esto supone un gran avance, ya que evita los problemas comentados anteriormente en el tratamiento de datos NGS, al no ser necesario un genoma ensamblado y permitiendo la detección de nuevas repeticiones (Novak *et al.*, 2013).

El funcionamiento de *RepeatExplorer* comienza con una comparación de todos los *reads* entre ellos, en busca de aquellos que presenten solapamiento por encima de un cierto valor umbral. Esto genera, para algunos paquetes de los analizados, más de 4500 billones de comparaciones y varios días de computación. Una vez se han comparado, se construye un grafo en el que los vértices o nodos son los *reads*, los puntos unidos por líneas representan los *reads* que solapan y la longitud de estas líneas manifiesta la similitud entre dichos *reads* (Novak *et al.*, 2010). De este modo, el análisis informático final de las secuencias se realiza con la función "*RepeatExplorer 2 clustering"*, en la que se ajustan los parámetros de análisis y se obtienen los resultados en forma de grafos y tablas que han de ser manualmente evaluados y analizados. Los resultados que se obtienen por *RepeatExplorer* son abundantes y complejos.

A continuación, se definen los parámetros y conceptos de mayor importancia para el tratamiento y la gestión de los resultados:

- **Monómeros:** Se denominan monómeros las diferentes secuencias contenidas dentro de una cadena de ADN consenso.
- **Proporción del genoma:** Calcula el porcentaje del genoma (con los datos de *reads* de los que dispone) que representa cada repetición encontrada (*cluster* de *reads*). Se calcula como el total de lecturas presentes en el *cluster* en cuestión, divididas entre el total de lecturas introducidas en el análisis, multiplicado esto por cien para calcular el porcentaje.
- **Tamaño:** Número total de lecturas en un *cluster*.
- **Consenso:** Representa la secuencia más probable de la repetición en cuestión, se obtiene tras el análisis del k-mero.
- **Longitud consenso:** Indica la longitud en número de bases de la secuencia consenso.
- **[V]:** Número de vértices del grafo.
- **[E]:** Número de aristas del grafo.
- **Longitud del K-mero:** Longitud del k-mero usado para la reconstrucción de la secuencia consenso.
- **Variante:** Identificador de la variante.
- **Calificación o *score*:** Medida del peso que tiene una variante en la construcción de la secuencia consenso.
- **Grafo:** En las diferentes variantes, se resaltan en color gris las partes del grafo usadas para la construcción del consenso.
- **Logo**: La secuencia de cada variante se muestra como un logo de ADN (gráfico de la secuencia) en el que la altura de las letras (que representan a las bases) muestran la proporción de lecturas que confirman esa posición.

3.5. Alineamiento e identificación de secuencias

Además de la plataforma *Galaxy-RepeatExplorer*, se han usado otros paquetes de *software* online para la obtención y el análisis de los resultados. Estos son la plataforma *MAFFT* (Kuraku *et al.*, 2013) y la herramienta *BLAST* del NCBI (Altschul *et al.*, 1997), las cuales se pueden usar a través de un servidor en línea; el programa *UGene* (Okonechnikov *et al.*, 2012); y el sitio web *UNIPROT* (Uniprot Consortium, 2016). La primera se usa para realizar el alineamiento de dos o más secuencias de nucleótidos. De esta manera, el alineamiento obtenido se analiza y edita con el programa *UGene*. Dicho programa permite obtener una secuencia consenso del alineamiento de los diferentes *contigs* que componen un *cluster*. Estas secuencias consenso son analizadas mediante el algoritmo *BLAST* a fin de identificar si dicha secuencia tiene homología con alguna

de las presentes en la base de datos del NCBI. Por último, el sitio web *UNIPROT* se usa como base de datos para conocer las características de diferentes proteínas que pueden tener relación con algunos de los datos obtenidos.

4. RESULTADOS

En el presente trabajo, se han usado los paquetes de datos indicados en la tabla 1 correspondientes a los *reads* del genoma transcrito de *S. senegalensis* en diferentes condiciones y órganos: etapas S0 del desarrollo larvario (etapa premetamórfica, periodo que comprende en torno a los primeros catorce días tras la eclosión) y S4 (etapa final de la metamorfosis donde tienen lugar los últimos cambios en la larva, generalmente se alcanza a partir de los veinte días después de la eclosión) de la metamorfosis del lenguado durante su desarrollo larvario (Klaren *et al.*, 2008); dos paquetes de la secuenciación del transcriptoma de las gónadas y cinco paquetes derivados del sistema inmune.

Cada uno de los paquetes de datos se envió mediante protocolo FTP (Protocolo de transferencia de archivos) a la plataforma de manera independiente. Una vez cargados en la plataforma, en el caso de los paquetes que se encontraban separados en varios sub-paquetes (diferentes librerías genómicas con códigos SRR, Tabla 1), como en el caso de muestras de gónadas y sistemas inmune, se realizó la concatenación de éstos para así tener un único paquete de datos. Por esta razón a partir de ahora se hará referencia a un único paquete de datos por condición u órgano.

El primer tratamiento de datos consiste en el análisis de la calidad de las secuencias mediante el programa *FastQC*. En el anexo se adjuntan las gráficas de calidad de secuenciación por bases y calidad de los cuatro paquetes de datos (Anexo AI). Aquí se adjunta un ejemplo de las gráficas de valores de calidad por base, obtenidas en el análisis de calidad del paquete de datos de la etapa S0 del desarrollo larvario y del paquete del tejido gonadal a fin de comparar las tecnologías Illumina y 454 respectivamente (Figuras 5 y 6).

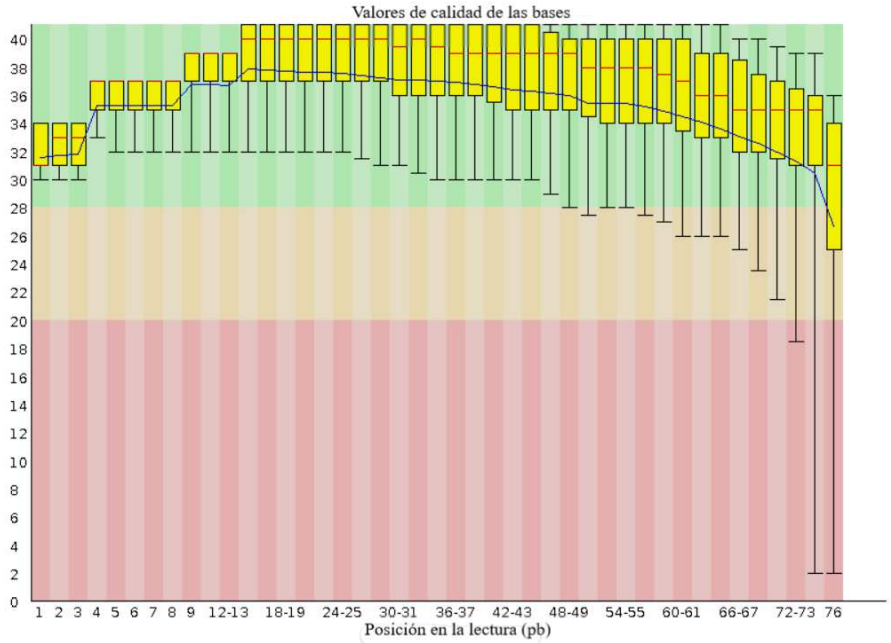


Figura 5. Distribución de los valores de calidad por base del paquete de datos del transcriptoma S0.

En el caso del paquete de S0, secuenciado por Illumina, se observa cómo la secuenciación es de una elevada calidad. En estas lecturas, la práctica totalidad de las posiciones se encuentran secuenciadas con un *Phred* superior a 30 (más de un 99,9% de fiabilidad de las lecturas), un valor excepcionalmente alto y que aporta una gran fiabilidad al método. También se ve reflejado cómo esta tecnología se ve limitada a la generación de *reads* de corta longitud, en este caso se han secuenciado lecturas de máximo 80 pares de bases (pb).

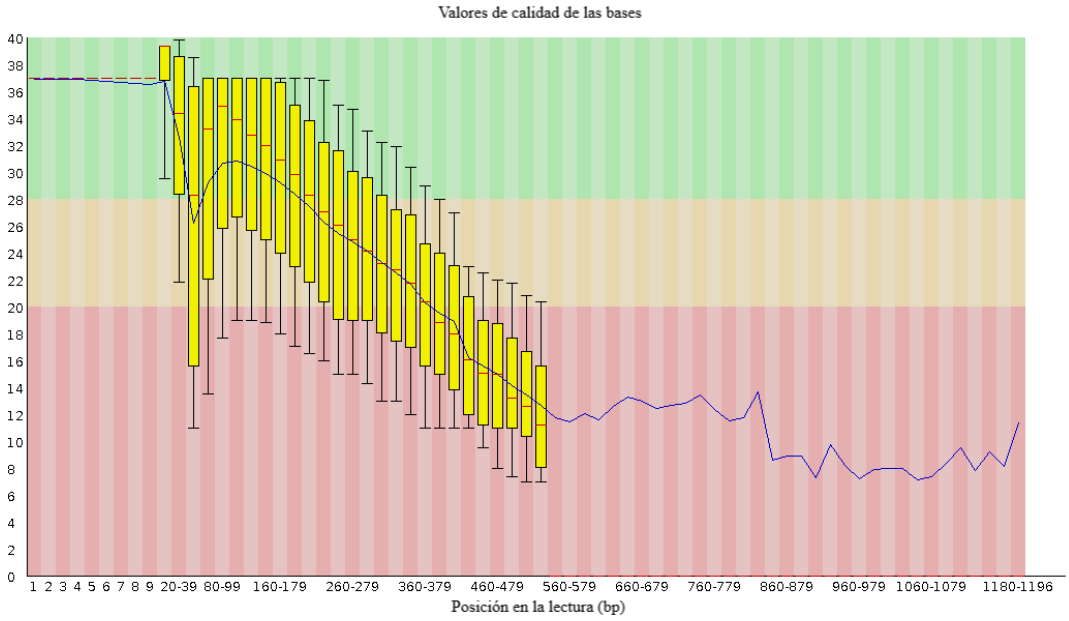


Figura 6. Distribución de los valores de calidad por base del paquete de datos del tejido gonadal.

Por el contrario, la secuenciación por 454, en este caso del transcriptoma del tejido gonadal, se ve caracterizada por una calidad más pobre, si bien la longitud de las lecturas es más

de cinco veces mayor a la conseguida con Illumina (las mejores calidades se encuentran en el rango 0-460 pb). Este mayor tamaño de *reads* es la principal fuerza de esta técnica de secuenciación de nueva generación (NGS), lo que le permite la detección de elementos repetitivos de mayor tamaño en el genoma. Los resultados de calidad descritos por cada plataforma de secuenciación se repiten en todos los lotes o paquetes de datos analizados.

Con la información obtenida, se repite varias veces en cada paquete el proceso de filtrado de los *reads*, con el objetivo de optimizar los filtros hasta obtener una buena cantidad de secuencias con una calidad adecuada. Los valores finales que se usan en cada parámetro del filtrado se adjuntan en la Tabla 2. Los parámetros finales de filtrado han permitido alcanzar un balance entre obtener secuencias lo más largas posibles y con una calidad adecuada. En la Tabla 3 se indica el número de lecturas que queda en cada paquete tras la concatenación y el filtrado de las secuencias, así como el muestreo final de lecturas que se realiza.

Tabla 2. Parámetros ajustados para el filtrado de las secuencias.

Paquete de datos tratado	**Muestreo**	**Límite mínimo de calidad (*phred*)**	**Porcentaje por encima del límite de calidad**	**Rango óptimo de las secuencias (Posición pb)**	**Máximo de bases sin secuenciar (Ns)**
Metamorfosis S0	No	20	95%	15-76	0
Metamorfosis S4	No	23	95%	12-70	0
Gónadas	No	15	95%	30-350	0
Sistema inmune	No	15	95%	35-240	0

Tabla 3. Resumen de los reads presentes en cada paquete de datos antes y después del proceso de filtrado.

Paquete de datos	***Reads* iniciales**	***Reads* tras el filtrado**	***Reads* muestreados en el análisis**
Metamorfosis S0	67.844.392	54.838.950	54.000.000
Metamorfosis S4	63.604.222	51.654.246	51.000.000
Gónadas	1.396.264	119.079	119.000
Sistema inmune	1.228.101	251.459	250.000

Con los datos filtrados se ejecuta finalmente el programa principal de la plataforma bioinformática *RepeatExplorer* para la búsqueda de los elementos repetitivos presente en la parte del transcriptoma analizado.

De manera general, los resultados muestran un número muy elevado de *clusters* (2572), o grupos de *reads*, de los cuales se analizarán en profundidad 32, identificados tras el análisis de los grafos y sus secuencias como elementos repetitivos de diferente naturaleza (Tabla 4).

Tabla 4. Clusters identificados en cada paquete de datos con la herramienta RepeatExplorer 2.

Paquete de datos	**Número de *clusters* identificados**
Metamorfosis S0	644
Metamorfosis S4	606
Gónadas	670
Sistema inmune	652

A fin de ayudar a la comprensión de los resultados, en los archivos de salida obtenidos se van a diferenciar dos "niveles" o "categorías". Los *clusters* que se observan en el "nivel 1" son todos los construidos por la propia plataforma de análisis, con una secuencia de ADN consenso derivada del análisis de los *k-meros*. Además, de entre dichos *clusters*¸ algunos son identificados por el algoritmo como especialmente significativos de representar una repetición. Adicionalmente, las herramientas del programa permiten acceder a un análisis en mayor profundidad, aportando una mayor información acerca de la construcción de la secuencia consenso de estos *clusters*. Dicha información complementaria, el análisis de cada *k-mero*, se considerará de "nivel 2". A modo de resumen se puede establecer que el nivel 2 comprende los diferentes *clusters* y secuencias que usa el algoritmo para construir el *cluster* más probable de nivel 1. Además, es importante destacar que en aquellos *clusters* de nivel 1 que el programa considera menos probables de constituir una repetición, no se aporta la misma información que en los que sí son significativos según sus estándares. En estos casos, no se obtiene una secuencia consenso determinada directamente por la plataforma, sino que se obtienen las secuencias de los diferentes *contigs* o agrupaciones dentro de un mismo *cluster*. Por tanto, es necesario hacer uso de la herramienta de análisis genético *MAFFT* externa a la plataforma *RepeatExplorer*, a fin de alinear las secuencias y obtener una secuencia consenso de manera manual para cada uno de los *clusters*. Es importante destacar que este análisis manual de las secuencias, en principio no identificadas por el programa con ningún elemento ya descrito previamente en las bases de datos como elemento repetitivo por la ausencia de homología con los mismos, no hacen sino confirmar la escasez de datos referentes a secuencias repetidas en las

bases de datos actuales, las cuales minimizan el contenido en las mismas porque ralentizan los motores de búsqueda de dichas plataformas de datos genómicos. Por tanto, pueden ser consideradas repeticiones *de novo* no descritas previamente.

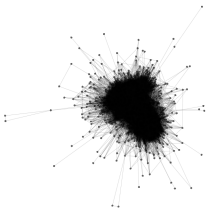
Por otra parte, se establecerá una nomenclatura que permita etiquetar fácilmente los *clusters* obtenidos, formulándose cada nombre con la abreviatura del paquete de datos al que pertenezca y un identificador *(Cl+paquete+Id)*.

A continuación, se exponen los resultados obtenidos en el análisis de cada uno de los paquetes de datos.

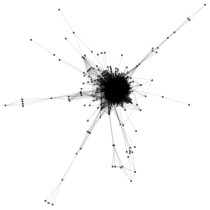
4.1. Paquete etapa S0 del desarrollo larvario.

A continuación, se exponen los resultados obtenidos tras el análisis de la secuenciación con Illumina del transcriptoma de la etapa S0 del desarrollo larvario. En este caso, se obtienen un total de 644 *clusters* de nivel 1. Se han seleccionado los más significativos de entre todos estos para el análisis manual de las repeticiones (Figura 7; Tabla 5).

ClS0_1

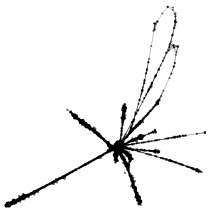


ClS0_2

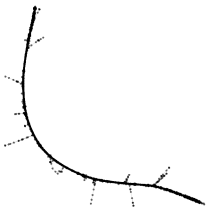


ClS0_3

ClS0_4



ClS0_5



ClS0_6

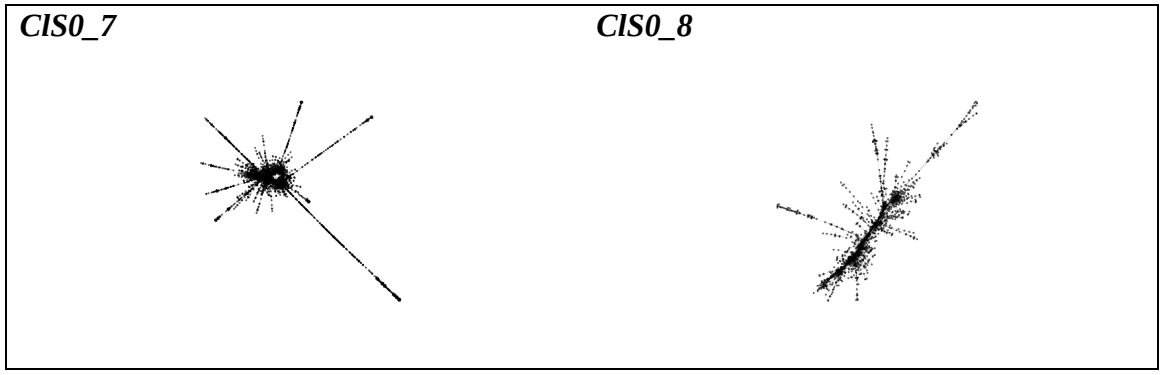


Figura 7. Grafos de los clusters más significativos encontrados en las lecturas del paquete S0.

Como se puede observar, los *clusters ClS0_1*, *ClS0_2*, *ClS0_7* y *ClS0_8* se caracterizan por tener una morfología estrellada. Por el contrario, los *clusters* restantes, cuya forma de "lazo" queda definida principalmente por una curva uniforme, carecen de esta multitud de caminos que divergen de una línea central.

Tabla 5. Características y parámetros que definen los clusters más significativos del paquete S0.

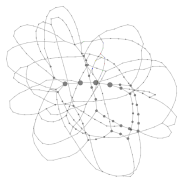
Cluster	**Proporción del genoma**	**Tamaño (N reads)**	**Longitud consenso (pb)**	**[V]**	**[E]**	**Clase de repetición**
ClS0_1	0.037	786	38	786	42.300	Satélite
ClS0_2	0.021	445	21	445	18.600	Satélite
ClS0_3	0.460	9747	490	9.744	3.780.000	Transposón
ClS0_4	0.390	8434	523	8.434	1.290.000	Transposón
ClS0_5	0.350	7410	2674	7.410	576.000	Transposón
ClS0_6	0.210	4437	105	4.437	375.000	Transposón
ClS0_7	0.160	3368	-	3.272	46.700	Satélite
ClS0_8	0.063	1357	-	1.357	12.100	Satélite

RepeatExplorer considera los *clusters ClS0_1*, *ClS0_2*, *ClS0_7* y *ClS0_8* posibles repeticiones tipo satélites. Esto encaja con el análisis visual de los grafos obtenidos, observándose las formas estrelladas características de los satélites. Por otra parte, se observan las formas de lazo de los transposones. Además, la longitud en pares de bases de los transposones es típicamente mayor que la de los satélites, lo que en este caso se cumple, observándose cómo los elementos transponibles tienen una secuencia, como mínimo, el doble de larga que el mayor de los satélites, llegando a ser hasta setenta veces más grande. Cabe destacar también que la proporción del genoma se calcula como el número de *reads* que conforman el *cluster* dividido entre el total de *reads* usados en el análisis multiplicado por cien, por lo que era de esperar que se obtuvieran valores relativamente bajos.

Por otra parte, en este caso, el algoritmo considera los *clusters ClS0_1* y *ClS0_2* como muy probables de constituir un satélite, por lo que aporta por sí mismo una secuencia consenso de dichos *clusters*, así como un análisis del k-mero y toda la información de nivel 2 (Anexo AIII.I). Como se ha explicado, los demás carecen de una secuencia consenso determinada por el programa, por lo que se ha hecho uso de *MAFFT* para alinear las secuencias de los *contigs* que componen cada uno de los *clusters* seleccionados. No obstante, los *clusters ClS0_7* y *ClS0_8* han tenido que ser descartados dado que no se podía obtener un alineamiento adecuado de las secuencias de los *contigs*.

Particularmente, los *clusters ClS0_1* y *ClS0_2*, como se ha comentado, son considerados por la plataforma como altamente significativos, por lo que se obtiene de ellos información de nivel 2. Por una parte, hay que considerar el enorme volumen de datos presentes en el análisis de la construcción de los *k-meros* (información de nivel 2), del orden de 50 *k-meros* para cada uno de los *clusters* que la plataforma trata como especialmente significativos, un total unos 500 *k-meros* (dos *clusters* significativos en este paquete y ocho en el caso de S4). Esto, sumado a que su importancia queda relegada a un segundo plano dado que los *clusters* definitivos que se llegan a construir a partir de estos *k-meros* son los que se exponían inicialmente (Figura 7). Se expondrá aquí una muestra explicativa de estos datos de nivel 2 (Tabla 6), mientras que los *k-meros* más significativos de cada *cluster* podrán encontrarse en el Anexo AIII.I.

Tabla 6. Información del k-mero ClS0_1_1 ClS0_2_1 como muestra de la información de nivel 2 adjunta en el anexo.

Variante	***K-meros***	***Score***	**Longitud del monómero**	**Grafo**
ClS0_1_1	15	0.4871	38	
CLS0_2_1	11	0.6304	21	

La información aportada por esta tabla indica que se trata de los *k-meros* identificados como *ClS0_1_1* y *ClS0_2_1*, uno de los muchos *k-meros* que es usado por el algoritmo para

llegar a construir el *cluster ClS0_1*. La columna "*k-mero"* indica el número de veces que se repite la secuencia monomérica, en este caso, la secuencia (ésta se encuentra en Anexos indicada bajo la tabla correspondiente, dado que su tamaño impedía incorporarla en el texto principal de la presente memoria) se haya repetida un total de quince veces para *ClS0_1_1* y once veces para *ClS0_2_1*. Por su parte, el valor del *score* informa acerca del peso que tiene la secuencia del *k-mero* en cuestión sobre la construcción del *cluster* consenso final. Finalmente, el *logo* (Figura 8) informa acerca de los nucleótidos que se encuentran en cada una de las posiciones de la secuencia del *k-mero*, así como el nivel de conservación de cada uno de ellos tras la comparación de las secuencias que dan lugar al mismo.

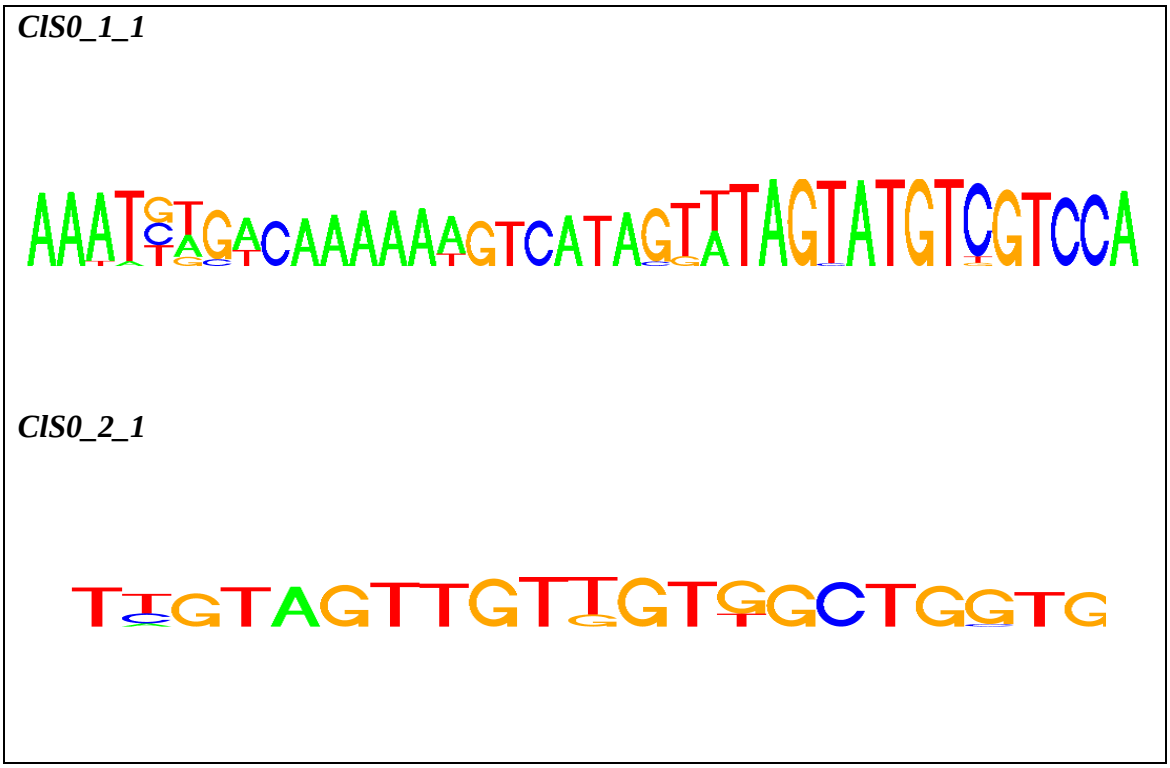


Figura 8. Logos de los k-meros comentados, la altura de las letras da información sobre la conservación de las posiciones.

Finalmente, las secuencias consenso obtenidas (Anexo AII.I) se analizan con *BLAST* para identificar posibles similitudes con la base de datos, teniendo en cuenta que los *clusters ClS0_7* y *ClS0_8* no han podido ser comparados en *BLAST* al no disponer de su secuencia. De todos los demás, se ha observado una única similitud significativa, la de la secuencia del *cluster ClS0_6* con "*Leucine-rich repeat-containing protein 3-like"*. A priori, puede entenderse que el algoritmo de *RepeatExplorer* identifica este tipo de genes como secuencias repetitivas ya que contienen repeticiones que tras su traducción darán lugar a una secuencia de aminoácidos repetida multitud de veces.

4.2. Paquete etapa S4 del desarrollo larvario.

Se realiza el procedimiento descrito anteriormente para el paquete de datos de la etapa S4 del desarrollo larvario, del transcriptoma secuenciado por la tecnología Illumina, y se obtienen un total de 606 *clusters* de nivel 1. De entre todos estos, los ocho que se exponen a continuación son los más significativos (Figura 9; Tabla 7).

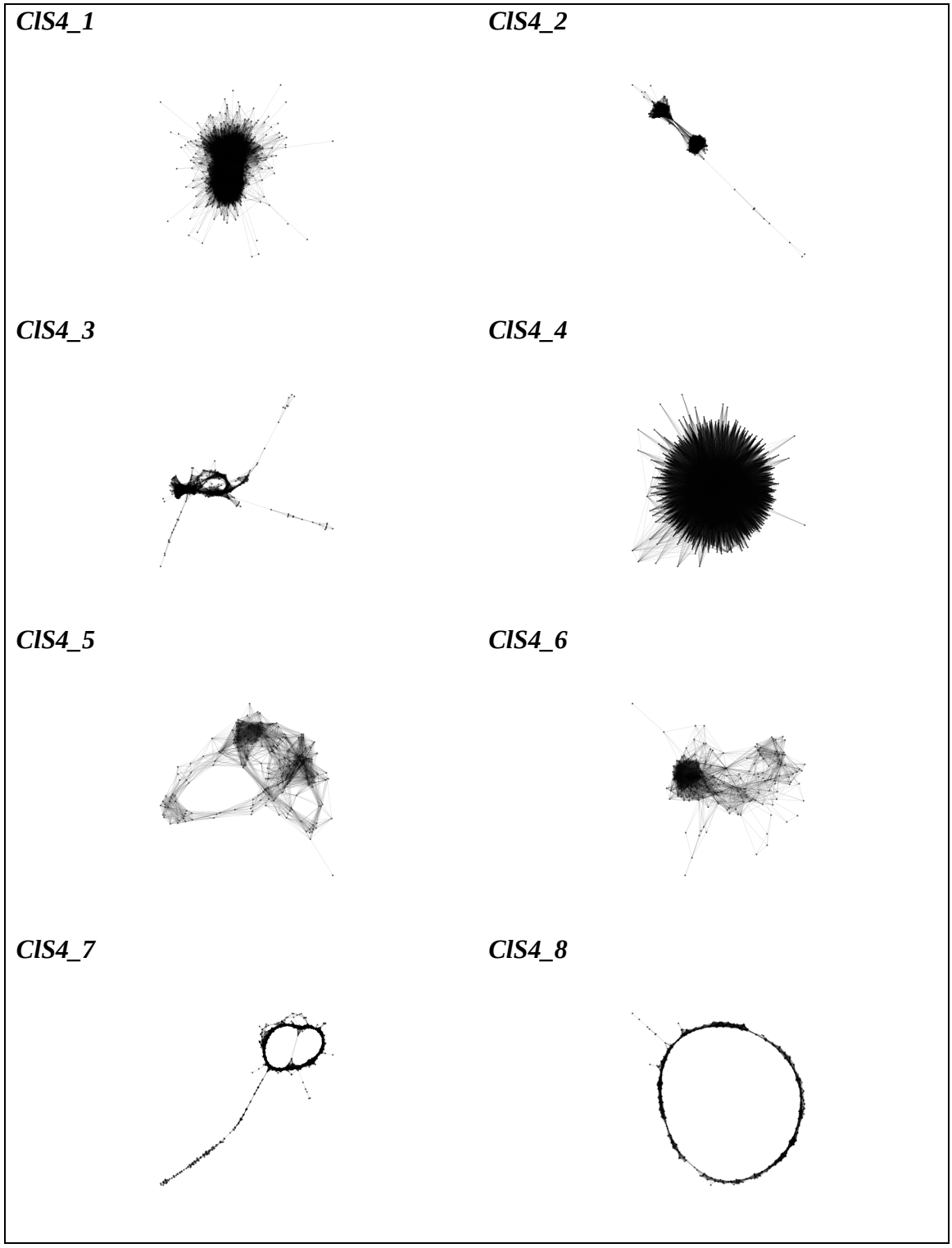


Figura 9. Grafos de los clusters más significativos encontrados en las lecturas del paquete S4.

Gráficamente se observa cómo los *clusters ClS4_1*, *ClS4_2*, *ClS4_3*, *ClS4_4*, *ClS4_5* y *ClS4_6* se caracterizan por ser más compactos, todos ellos poseen una región con un alto índice de solapamiento donde se unen una gran cantidad de *reads*. Por otra parte, los *clusters ClS4_7* y *ClS4_8* se caracterizan por tener una forma de lazo o circunferencia bien definida.

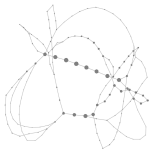
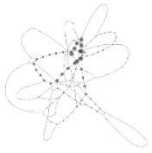
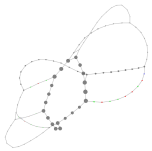
Tabla 7. Características y parámetros que definen los clusters más significativos del paquete S4.

Cluster	**Proporción del genoma**	**Tamaño (N reads)**	**Longitud consenso (pb)**	**[V]**	**[E]**	**Clase de repetición**
ClS4_1	0.038	751	38	749	52.200	Satélite
ClS4_2	0.027	529	27	529	41.200	Satélite
ClS4_3	0.026	514	138	514	18.500	Satélite
ClS4_4	0.017	337	42	337	18.700	Satélite
ClS4_5	0.012	229	102	229	4.190	Satélite
ClS4_6	0.011	218	57	218	4.330	Satélite
ClS4_7	0.071	1406	702	1.406	47.700	LTR
ClS4_8	0.036	722	1160	722	11.600	LTR

La plataforma identifica los *clusters ClS4_7* y *ClS4_8* como posibles transposones, concretamente LTR, mientras que todos los demás son considerados posibles ADN satélite. Esto se contrasta manualmente con el estudio de los grafos (Fig. 9) y los parámetros asociados a cada uno de ellos (Tabla 7). Así, las repeticiones con una longitud de secuencia varias veces superior a la longitud media de las lecturas, como es el caso de los LTR, 702 y 1160 pares de bases (pb) (*ClS4_7* y *ClS4_8* respectivamente), da lugar a la forma de lazo observada. En contraposición, secuencias más cortas, que no lleguen a alcanzar el tamaño medio de los *reads*, presentan una característica forma estrellada en el grafo, lo que se observa claramente, por ejemplo, en el caso del *cluster ClS4_1*.

En este paquete, los ocho *clusters* seleccionados disponen por parte del programa de la información de nivel 2. Tal y como se hizo anteriormente, aquí únicamente se expone el *k-mero* más influyente en la construcción del cada uno de los *clusters* definitivos (Tabla 8), en el anexo se adjuntan los cinco *k-meros* más influyentes en la construcción de los *clusters* (Anexo AIII.II).

Tabla 8. Información de los k-meros con mayor score para cada uno de los clusters seleccionados en el paquete S4 como muestra de la información de nivel 2 adjunta en el anexo.

Variante	***K-meros***	***Score***	**Longitud del monómero**	**Grafo**
ClS4_1_1	11	0.5639	38	
CLS4_2_1	15	0.412	27	
CLS4_3_1	11	0.454	138	
CLS4_4_1	15	0.72	42	
CLS4_5_1	15	0.578	102	
CLS4_6_1	15	0.45	57	
CLS4_7_1	19	0.561	702	
CLS4_8_1	15	0.976	1164	

Al igual que en el paquete de datos del desarrollo larvario S0, se obtienen tablas con los mismos parámetros. Es interesante ver cómo a pesar de tratarse de los *k-meros* con un mayor *score*, la mayoría por encima de 0'5 o muy cerca de este valor, el algoritmo aún tiene en cuenta multitud de variaciones hasta conformar el *cluster* consenso. Esto es algo que se aprecia muy bien gráficamente en cómo cambia la morfología de los grafos entre los *clusters* consenso y los *k-meros*, pese a que las secuencias consenso y las secuencias de estos *k-meros* seleccionados coincidan en su mayoría.

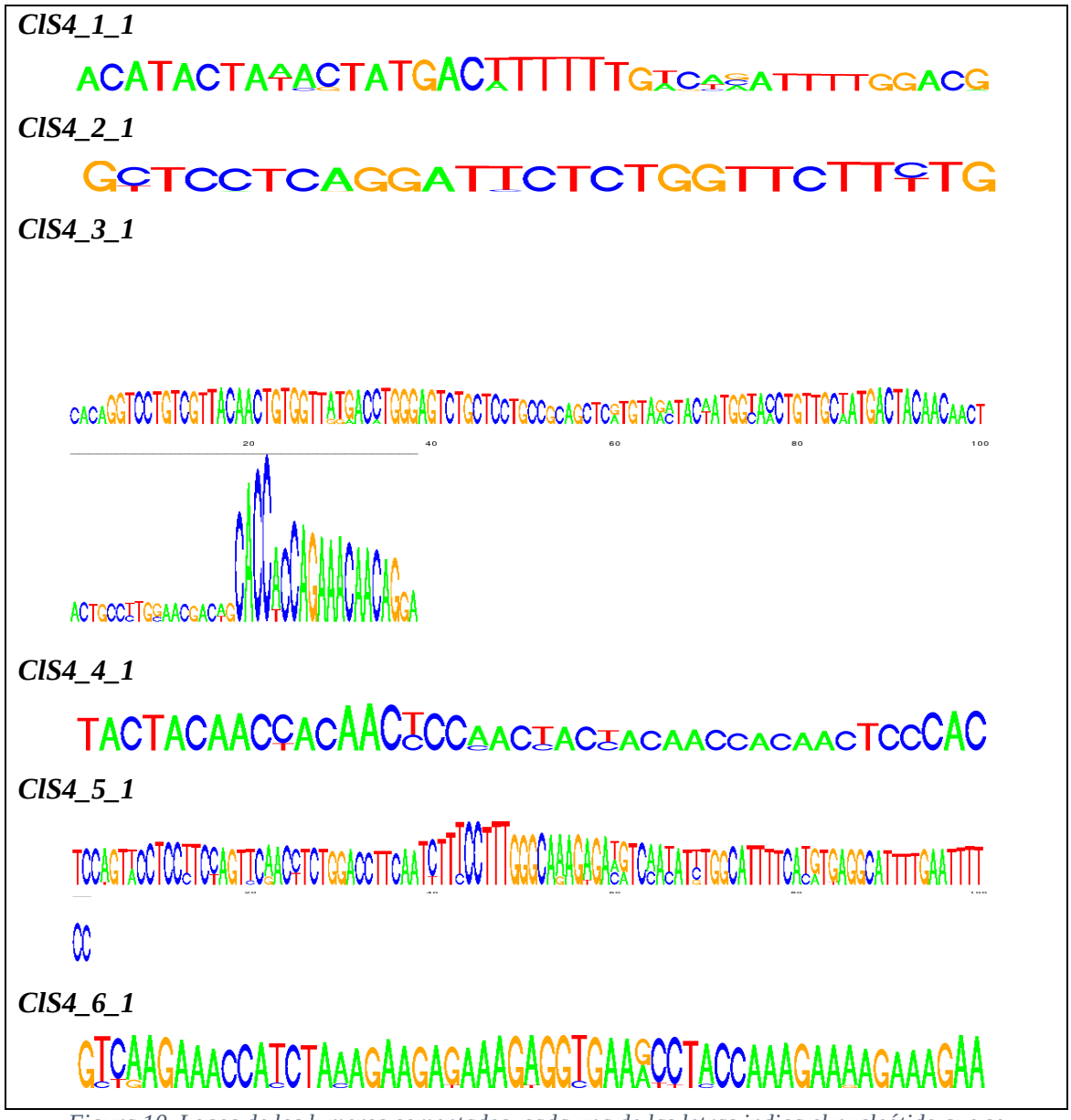


Figura 10. Logos de los k-meros comentados, cada una de las letras indica el nucleótido que se encuentra en su posición. Los logos de los k-meros ClS4_6_7 y ClS4_6_8 se adjuntan en el anexo dado que la longitud de su secuencia era inabarcable en el espacio aquí disponible.

En cuanto a los *logos*, permiten detectar visualmente de una manera rápida cómo existen regiones con una cobertura mucho mayor que el resto de la secuencia, como ocurre en el

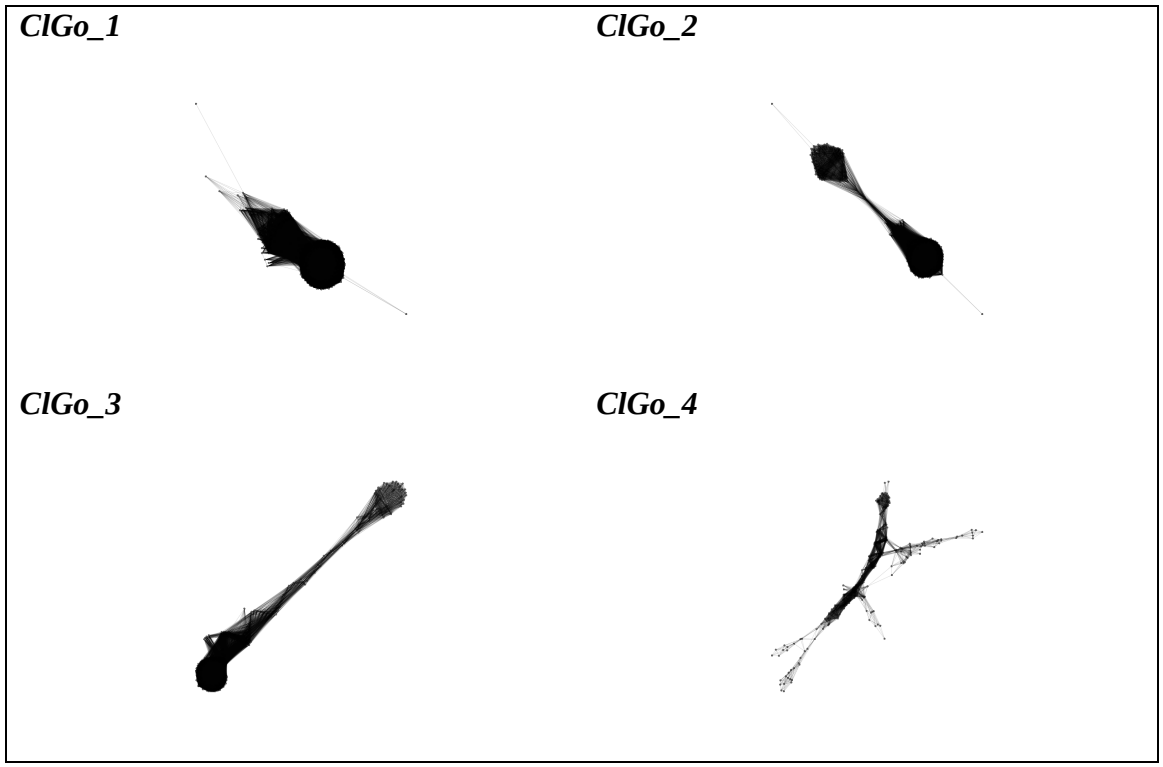
caso del *k-mero ClS4_3_1*, en el que se observa una zona muy conservada en C y A en el extremo final de la secuencia.

La secuencia de cada uno de los *clusters* (Anexo AII.II) se ha analizado en la plataforma BLAST del NCBI, a fin de identificar si alguna de las secuencias presentaba un buen valor de similitud con alguno de las secuencias alojadas en sus bases de datos.

La mayoría de las secuencias de los *clusters* no presentan homología con las depositadas en la base de datos NCBI. En aquellos casos en los que ha habido homología se puede tratar de secuencias que contienen elementos reconocidos por el programa como repetidos a lo largo del transcriptoma analizado, conteniendo dominios o secuencias satélites presentes en los diversos genes obtenidos. De esta manera, se ha detectado cómo los *clusters ClS4_5* y *ClS4_7* contienen una secuencia de ADN con homología a la nucleoproteína *AHNAK* y, en el caso de *ClS4_5*, también con periaxina. Por su parte, se observa cómo *ClS4_3* se relaciona con una proteína con actuación supresora tumoral (*deleted in malignant brain tumors 1 protein, DMBT1*). Por último, el *cluster ClS4_8* presenta homología con el gen *zonadhesina*, cuya función es la adhesión célula-célula de los gametos.

4.3. Paquete transcriptoma de las gónadas.

En tercer lugar, se expone el análisis de los datos de las secuencias obtenidas por 454 de los tejidos gonadales de *S. senegalensis*, obteniéndose de este paquete un total de 670 *clusters*. En la figura 11 se exponen los grafos de los *clusters* seleccionados.



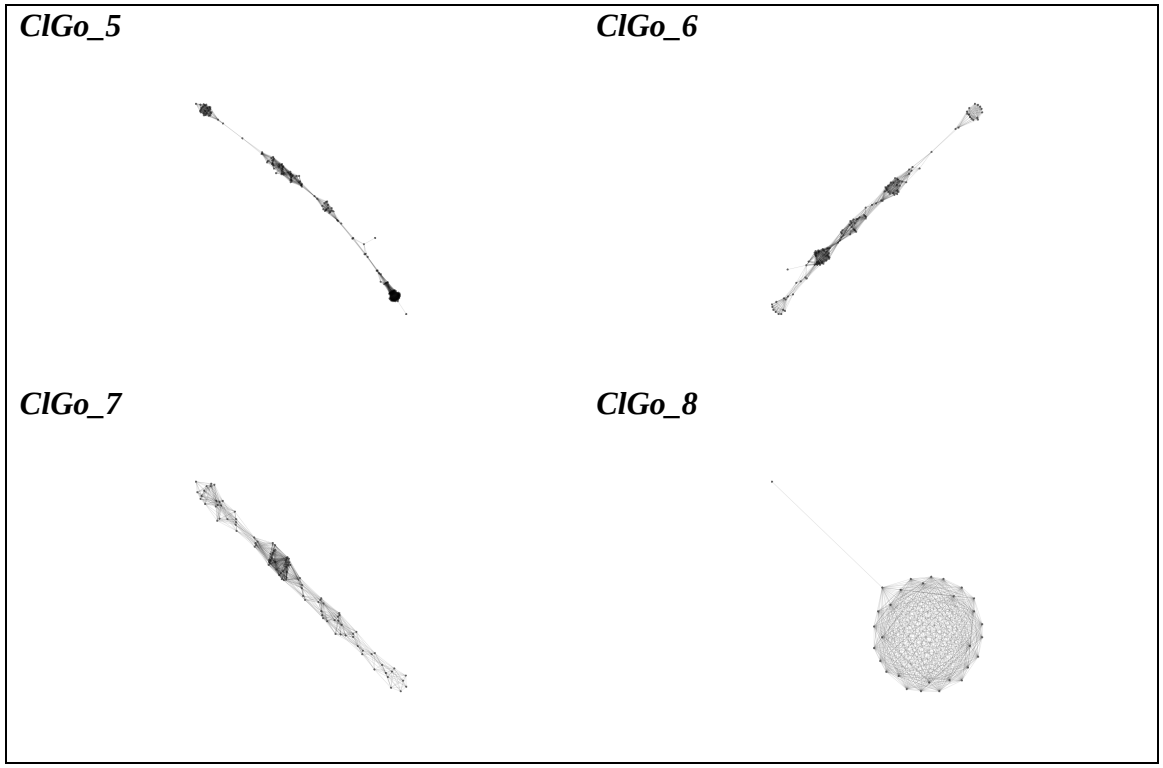


Figura 11. Grafos de los clusters más significativos encontrados en las lecturas del paquete de las gónadas.

En este paquete de datos los grafos no presentan una morfología tan evidente como en los paquetes secuenciados por Illumina, no obstante, sigue siendo posible distinguir la forma estrellada, una morfología más expandida y ensanchada en los *clusters ClGo_1*, *ClGo_2* y *ClGo_8*. En el resto de los grafos parece observarse una forma más cercana al lazo, como los obtenidos en el paquete del transcriptoma del desarrollo larvario en S0.

Tabla 9. Características y parámetros que definen los clusters más significativos del paquete de las gónadas.

Cluster	**Proporción del genoma**	**Tamaño (N reads)**	**Longitud consenso (pb)**	**[V]**	**[E]**	**Clase de repetición**
ClGo_1	0.7	838	261	838	281.000	Satélite
ClGo_2	0.31	364	840	364	41.800	Satélite
ClGo_3	0.3	364	1144	362	28.000	LTR
ClGo_4	0.25	300	1677	300	5.440	LTR
ClGo_5	0.13	152	2282	152	1.890	LTR
ClGo_6	0.1	120	1930	120	1.240	LTR
ClGo_7	0.085	101	1692	101	1.040	Satélite
ClGo_8	0.025	30	459	30	407	Satélite

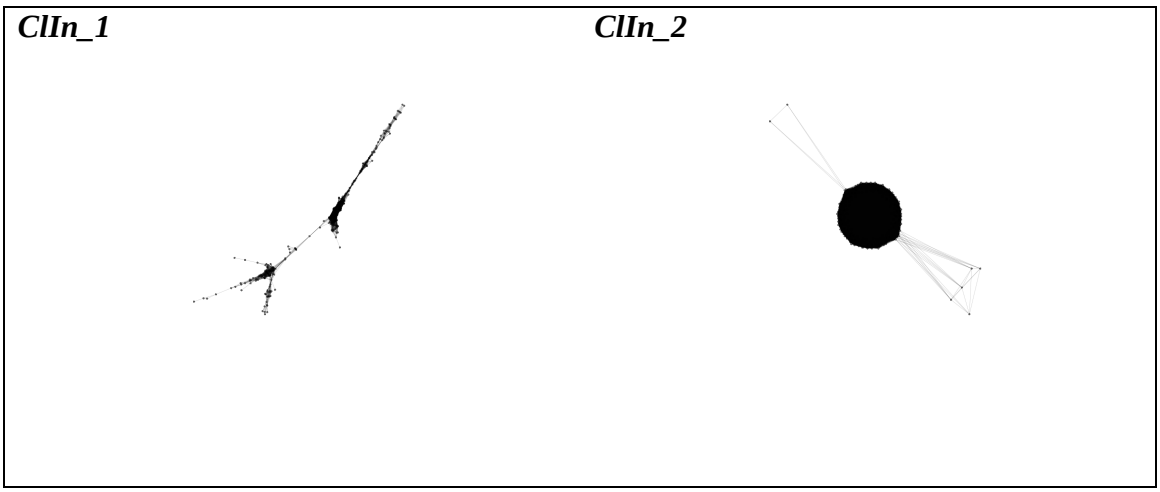
Con los valores de los diferentes parámetros (Tabla 9) se detecta cómo los *clusters ClGo_1*, *ClGo_2*, *ClGo_7* y *ClGo_8*, que en base al programa y a la observación de los grafos parecen tratarse de satélites, resultan tener una longitud mayor que la observada en este tipo de repeticiones en los dos paquetes de datos anteriores. Del mismo modo ocurre con los otros cuatro *clusters*, identificados visualmente como transposones y marcados por el programa como probables retrotransposones de tipo I. Se considera que este aumento de tamaño en los *clusters* formados se debe principalmente a la tecnología utilizada para la secuenciación. La capacidad de 454 de generar lecturas mucho más largas que Illumina permite la detección de elementos de mayor tamaño de manera más fácil.

En esta ocasión, puntualmente el programa indica cierta similitud de los *clusters ClGo_3*, *ClGo_4*, *ClGo_5* y *ClGo_6* con otros LTR anotados en bases de datos. Así, ha identificado cierta similitud en *ClG0_3* a un LTR tipo *DIRS-RT*; en *ClG0_4* se detectan indicios de LTR tipo *Ty3_gypsy:Ty3-INT*; en *ClG0_5* hay similitud con un LTR tipo *Ty1_copia:Ty1-RH*; y finalmente *ClG0_6* se identifica parcialmente con un LTR tipo *Ty1_copia:Ty1-INT*.

Tras el análisis de las secuencias con *BLAST* (Anexo AII.III) se encuentran dos secuencias con un alto grado de similitud, la del *cluster ClGo_2* y la del *ClGo_8*, la primera se identifica con *translation elongation factor 1 beta 2* (*eef1b2*) y la segunda con la proteína ribosómica *L39*. Las proteínas ribosomales se encuentran en multitud de copias en el genoma, por lo que el algoritmo identifica estas secuencias como secuencias repetidas presentes en los *reads* del transcriptoma.

4.4. Paquete transcriptoma del sistema inmune.

Finalmente, se analiza el paquete de datos de secuenciación por la tecnología 454 del transcriptoma relacionado con el sistema inmune, generándose un total de 652 *clusters*, de los cuales se han seleccionado los ocho siguientes para un análisis manual en mayor profundidad (Figura 12).



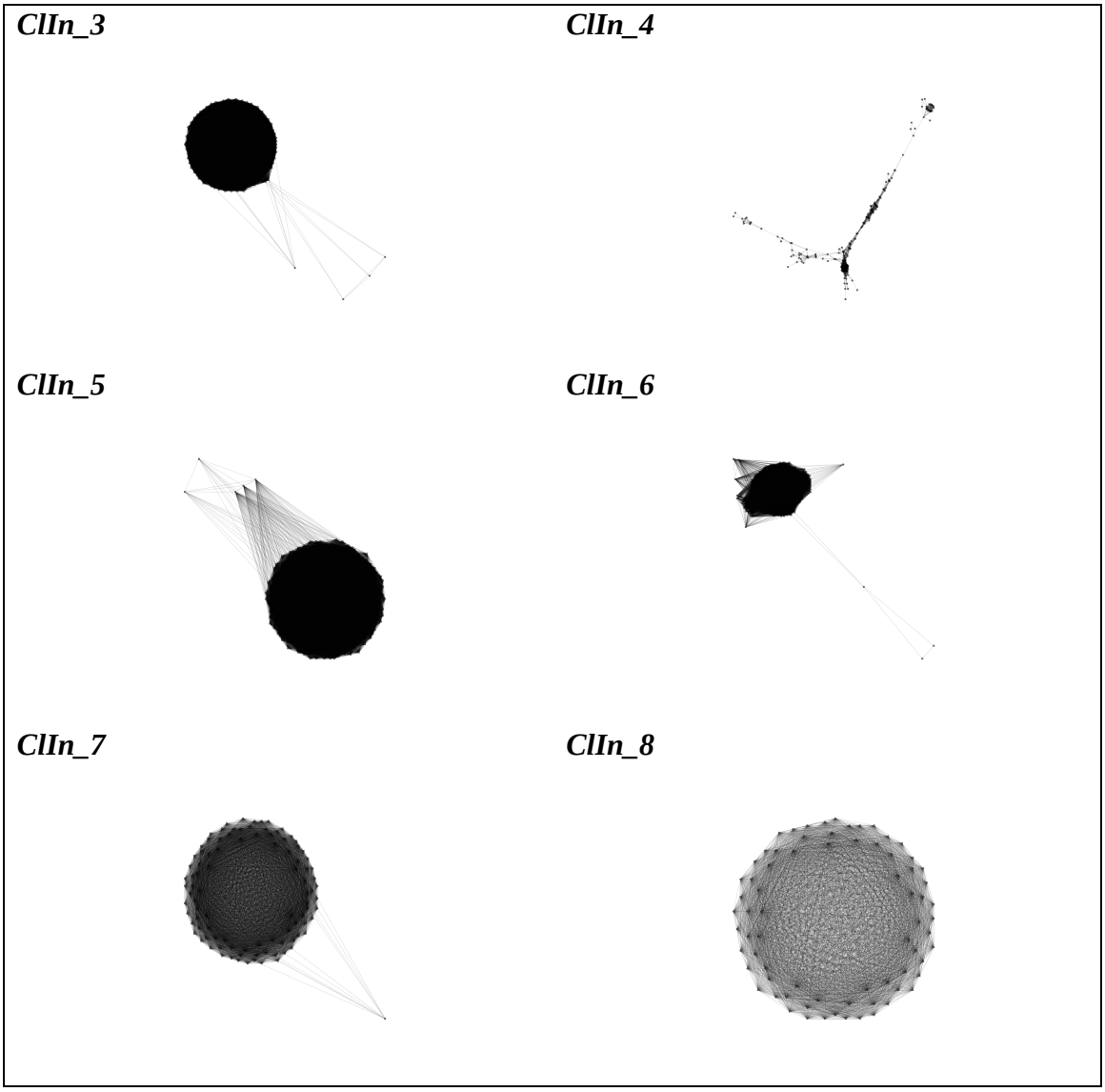


Figura 12. Grafos de los clusters más significativos encontrados en las lecturas del paquete del sistema inmune.

En este paquete los *clusters ClIn_1* y *ClIn_4* poseen la forma de lazo observada hasta ahora como morfología típica de los transposones. Por otra parte, los demás *clusters* presentan una morfología estrellada, con los *reads* unidos entre sí por un gran número de conexiones, observándose una alta densidad de solapamiento.

Los parámetros de cada uno de los *clusters* (Tabla 10) reflejan cómo la tecnología de secuenciación 454, tal y como se ha comentado anteriormente, influye en la longitud de los elementos detectados, siendo estos de mayor tamaño que los identificados con Illumina. Así mismo, el análisis de los grafos se ve respaldado por los valores de longitud observados, con los *clusters ClIn_1* y *ClIn_4*, que se consideraban LTR, de hasta cinco veces mayor longitud que los demás.

Tabla 10. Características y parámetros que definen los clusters más significativos del paquete del sistema inmune.

Cluster	**Proporción del genoma**	**Tamaño (N reads)**	**Longitud consenso (pb)**	**[V]**	**[E]**	**Clase de repetición**
ClIn_1	0.13	331	1056	331	6660	Transposón
ClIn_2	0.12	292	503	292	40.500	Satélite
ClIn_3	0.099	247	337	247	29.400	Satélite
ClIn_4	0.079	198	1076	198	1.950	Transposón
ClIn_5	0.078	196	235	196	18.300	Satélite
ClIn_6	0.077	193	196	193	16.300	Satélite
ClIn_7	0.04	99	299	99	4.760	Satélite
ClIn_8	0.029	73	301	73	2.620	Satélite

Con el análisis de las secuencias en *BLAST* se ha encontrado que los *clusters ClIn_3* y *ClIn_8* se identifican con genes que codifican para proteínas ribosómicas (*40S ribosomal protein S12* y *Ribosomal protein L27a* respectivamente) que como se ha argumentado, se encuentran muy repetidos y tienen una elevada expresión, por lo que el programa los detecta como secuencias repetitivas del transcriptoma. Por otra parte, el *cluster ClIn_1* se identifica con *tubulin alpha-1A chain* y *tubulin alpha-1B chain*; el *cluster ClIn_4* con *β-Tubulin*; y *ClIn_7* con *parvalbumin alpha* y *beta*.

4.5. Análisis comparativo de los resultados obtenidos en los paquetes de datos usados

Con todos estos resultados, se puede realizar una comparativa entre los diferentes paquetes de datos, con las diferentes condiciones bajo las que se obtuvo el transcriptoma y las tecnologías de secuenciación masiva 454 e Illumina.

A modo de resumen, se adjunta la tabla 11 donde se recopila qué representaba cada uno de los ocho *clusters* seleccionados en cada paquete en lo que se ha denominado nivel 1.

Tabla 11. Tabla resumen de los resultados de nivel 1.

Paquete de datos	**Tecnología**	***Clusters* transposones (N)**	***Clusters* satélites (N)**	***Otros clusters identificados* (N)**
Metamorfosis S0	*Illumina*	4	4	1
Metamorfosis S4	*Illumina*	2	6	4
Gónadas	*454*	4	4	2
Sistema inmune	*454*	2	6	4

Se ha obtenido una proporción similar de LTR y de satélites en los paquetes con las diferentes tecnologías, no solo en esta muestra de ocho *clusters* de cada uno de ellos, sino de manera general entre los más de 600 *clusters* obtenidos en los cuatro paquetes de datos. Sin embargo, sí que hay una diferencia fundamental entre los paquetes secuenciados por 454 e Illumina, y es que en los primeros se han identificado elementos repetitivos cuya secuencia consenso era varias veces más larga que en los de Illumina. Igualmente, la tecnología Illumina aporta la ventaja de obtener una cantidad enorme de lecturas, obteniéndose así grafos mucho más densos y robustos.

4.6. *Workflow* o diagrama de trabajo

A través de la plataforma *Galaxy-RepeatExplorer* se ha diseñado un diagrama de trabajo en el que se indican los pasos seguidos para llegar a obtener los resultados presentados en este trabajo (Figura 13).

En el diagrama se indica cómo en primer lugar se cargan los datos a la plataforma, tras lo cual se concatenan para trabajar con un único paquete de datos (uno para cada transcriptoma). Este paquete de datos se analiza con la función *FastQC*, que aporta la información necesaria para ejecutar el filtrado de los *reads*. Este filtrado se repite múltiples veces hasta alcanzar un balance adecuado en los filtros, que permita mantener la mayor cantidad de lecturas posibles con la mayor calidad que se pueda. Finalmente, se obtiene un nuevo paquete de datos filtrados que ha de ser analizado por el algoritmo de *RepeatExplorer 2 clustering*, obteniendo así los *clusters* de cada paquete junto con toda la información asociada a cada uno de los *clusters*.

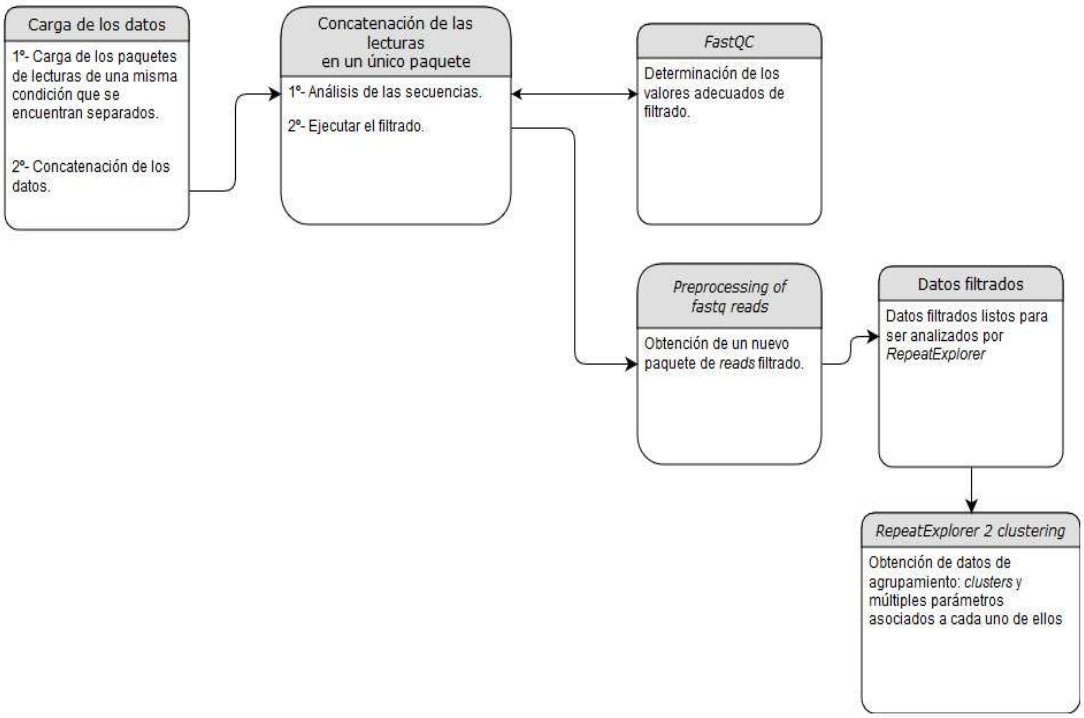


Figura 13. Diagrama del trabajo realizado con las diferentes funciones de la plataforma Galaxy-RepeatExplorer.

Con esto, se abren múltiples rutas de trabajo, comenzando con la selección de los *clusters* a analizar manualmente, se hace uso posteriormente de las múltiples plataformas y programas indicados a lo largo del trabajo a fin de tratar y clasificar la información obtenida.

5. DISCUSIÓN

Las repeticiones identificadas, al menos las relacionadas con los 32 *clusters* de nivel 1 analizados en profundidad en este trabajo, poseen las características y presentan los parámetros adecuados como para afirmar que se trata de elementos transponibles o satélites, según el caso. Sin embargo, la mayor parte de estas repeticiones no se encuentran anotadas, lo que hace patente la enorme brecha que existe en el conocimiento de este tipo de elementos. Además, es interesante destacar el hecho de que el uso de lecturas obtenidas a partir de transcriptomas limita la detección de elementos transponibles a los transposones de tipo I, los retrotransposones, ya que son los que hacen uso de la transcripción para movilizarse a lo largo del genoma. Si bien esto es una limitación en cuanto al tipo de repeticiones que se pueden hallar, también aporta una ventaja, y es que el volumen de datos en *reads* sigue siendo enorme, como se ha podido observar en este trabajo, pero al reducir el número de repeticiones a las que pueden transcribirse, se facilita el tratamiento de los datos, enfocando el estudio a un número más reducido de clases de elementos repetitivos.

En cuanto a los genes con los que existía una similitud en la secuencia según la plataforma *BLAST*, se pueden extraer varias conclusiones. Es posible que las partes de la secuencia que

identifica como homólogas entre el gen y el *cluster* sean regiones conservadas, compartidas entre el gen y el elemento repetitivo. Por otro lado, genes que codifican para proteínas ricas en un aminoácido pueden estar siendo detectados por el programa como un posible LTR, basado en la presencia de las secuencias repetidas que darán lugar a esos aminoácidos. Igualmente, hay que destacar que el transcriptoma se secuencia cuando el ARN transcrito por la célula ha pasado tras un proceso de maduración, en el que los intrones han sido eliminados y solo quedan los exones y las regiones UTR en los extremos, de manera que se descarta la posibilidad de que los elementos repetitivos hallen su similitud con los intrones de los genes identificados.

En lo que a los LTR respecta, se identificaron varios de ellos como parcialmente similares a otros ya anotados. En primer lugar, se observa cómo el *cluster ClG0_3* tiene cierta similitud con un LTR tipo DIRST-RT, siendo este un tipo de retrotransposones englobados filogenéticamente en la misma familia que los de tipo *gypsy* (Jurka *et al.*, 2007). Estos últimos, los *gypsy*, se encuentran parcialmente presentes en los *clusters ClG0_4*, *ClG05* y *ClG0_6*, el primero similar a un *gypsy Ty3* mientras que los otros dos son similares a un *gypsy Ty1_copia*. Estos elementos transponibles suponen una familia de LTR de gran importancia, habiéndose demostrado cómo diferentes genes que secuencian para proteínas críticas en procesos de desarrollo embrionario y de defensa del sistema inmune, se han desarrollado a partir de elementos tipo *gypsy* (Ono *et al.*, 2001; Volff *et al*., 2001; Jurka *et al.*, 2007; Serrato-Capuchina y Matute, 2018).

Como se ha explicado, los elementos transponibles ejercen un gran impacto en la evolución de las especies, no solo con su capacidad de originar nuevos genes, sino también porque, dada su naturaleza de elementos móviles, a menudo provocan con su inserción la alteración de regiones codificantes o reguladoras de los genes presentes en el genoma, pudiendo afectar negativamente a su expresión y funcionalidad (Oliver y Greene, 2009; Feschotte, 2008). Del mismo modo, los elementos transponibles pueden provocar expansiones genómicas y generar nuevas variantes cromosómicas a consecuencia de inversiones en el material genético. Todos estos cambios, que pueden ser inducidos muy rápidamente (Van de Lagemaat *et al.*, 2003; Flutre *et al.*, 2011) y en respuesta a factores abióticos, pueden proporcionar nuevas características fenotípicas sobre las que la selección puede actuar (Hollister y Gaut, 2009; Arkhipova y Meselson, 2005). Por todo esto, a consecuencia de su capacidad para aportar variabilidad, se ha planteado la hipótesis de que los elementos transponibles se mantienen en los genomas a causa de la selección (Vinogradov, 2004; Fablet y Vieira, 2011). Además de esto, estudios recientes destacan la capacidad de los elementos transponibles para llevar a cabo una transferencia horizontal, entre genomas de diferentes especies e individuos, gracias a diferentes vectores como virus y parásitos. Este tipo de movilidad puede haber sido el origen de elementos como el *LINE-L1*, el elemento transponible más extendido en mamíferos, lo que pone de

manifiesto la importancia de este fenómeno y el impacto de los elementos transponibles en los cambios genómicos (Ivancevic *et al.,* 2018). En este sentido, los datos aportados en este trabajo abren nuevas posibilidades de estudio del papel de estos elementos transponibles en etapas del ciclo vital tan importantes como la metamorfosis o la determinación sexual.

Por otra parte, el estudio de los elementos transponibles no debe restar importancia a la presencia de satélites. Siendo la principal función de estas secuencias repetidas la organización y regulación de regiones centroméricas, aún se desconoce mucha información acerca de otras funciones que pudiesen efectuar. Actualmente, el estudio de secuencias de ADN satélite se ha visto enormemente potenciado gracias a herramientas como *RepeatExplorer*, con la que grupos de investigación tratan de desarrollar el denominado "*satelitoma"*, el conjunto de todos los *clusters* de ADN satélite presentes en el genoma de un organismo (Ruiz-Ruano *et al.,* 2016). Estos satélites detectados son elementos de gran utilidad, ya que los polimorfismos presentes en su secuencia pueden ser usados como marcadores con los que establecer relaciones parentales entre diferentes individuos (Shah *et al.,* 2016). En estudios recientes se han realizado técnicas como la *Fluorescence in situ Hybridization* (*FISH*), que han permitido mapear diferentes satélites sobre cromosomas, advirtiéndose así diferentes tipos de organización: satélites que se organizan formando agrupaciones dentro de regiones específicas de los cromosomas, otros que no hacen forman estas agrupaciones y una mezcla de ambos (Ruiz-Ruano *et al.,* 2016). Todas estas organizaciones espaciales y variaciones en la secuencia hacen de los satélites un potente tipo de marcador molecular.

6. PERSPECTIVAS FUTURAS

En el presente trabajo se han analizado con herramientas bioinformáticas 8 paquetes de una base de datos de lecturas de secuenciación del transcriptoma de *S. senegalensis* y *S. solea*. Dicha base de datos contiene un total de 51 paquetes de *S. senegalensis*, y, por tanto, aún resta el estudio de otros 43 paquetes de esta especie. Esto se traduce en casi 1'7 x 10^9 *reads* que necesitan ser tratados. Del mismo modo, hay que tener en cuenta que se han podido analizar un total de 32 *clusters* de los más de 2500 obtenidos en este análisis. Adicionalmente quedarían por analizar los 43 paquetes de lecturas de la especie *S. solea* (Tecnología Illumina, con más de 2'1 x 10^9 *reads*) presentes en dicha base de datos, lo que permitiría hacer un análisis comparativo de elementos repetitivos entre dichos genomas transcritos.

Siendo evidente la labor necesaria con estos de datos, aún sería conveniente que, una vez analizado y caracterizado el transcriptoma completo de estas especies, donde se podría obtener una gran cantidad de información sobre las diferentes familias de elementos repetitivos transcritos en estos peces planos, se pasase a realizar esta metodología con el genoma completo

o la información que haya de ellos en bases de datos y genotecas tipo BACs, ampliando así aún más los tipos de elementos repetitivos que podrían encontrarse.

Por otra parte, una vez conocidas las secuencias de los elementos repetitivos obtenidos en el presente trabajo, se podría diseñar cebadores (*primers*) para realizar PCR sobre el genoma de esta especie y otras de peces planos para, en primera lugar identificar y analizar más profundamente la presencia, polimorfismos y evolución de los elementos encontrados y, posteriormente realizar un estudio citogenético mediante la técnica FISH, para localizar físicamente estas secuencias en los cromosomas de esta especie, pudiendo así determinar su perfil de distribución así como para llevar a cabo diferentes análisis comparativos con otros genomas conocidos mediante el estudio de su sintenia.

7. CONCLUSIONES

- Tras el análisis de lecturas de ADN del transcriptoma de la especie de lenguado *S. senegalensis*, se han encontrado un total de 2.572 secuencias (*clusters)* de elementos repetitivos tipo satélite y elementos transponibles en los paquetes de lecturas analizados mediante la plataforma *Galaxy-RepeatExplorer*.
- En los paquetes secuenciados por Illumina se han hallado 6 *clusters* identificados como posibles transposones, de los cuales 2 son LTR, y otros 10 *clusters* identificados como satélites.
- En los paquetes secuenciados por 454 se han encontrado 6 *clusters* que pueden constituir elementos transponibles, de los cuales concretamente hay 4 posibles LTR. Además, se han identificado otros 10 *clusters* como posibles satélites.
- Se observa que la mayoría de las repeticiones halladas mediante estos *clusters* no se encuentra anotada en las bases de datos consultadas.
- Se confirma la potencialidad de estas herramientas bioinformáticas de alto rendimiento para detectar elementos repetitivos *de novo* a partir de lecturas de genomas no ensamblados.
- Tras el análisis de secuencias procedentes de dos tipos diferentes de tecnologías de secuenciación de nueva generación (NGS) como son Illumina y 454, se hace patente la versatilidad que aporta a la metodología de análisis realizada el uso de estas tecnologías que generan lecturas de diferente longitud.
- Se establece un *workflow* que permite sintetizar los pasos realizados durante el uso de la plataforma *Galaxy-RepeatExplorer*, lo que permite extender esta metodología de análisis a otros paquetes de datos.

8. BIBLIOGRAFÍA

- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), 3389-3402.
- APROMAR (2017). La acuicultura en España. *Asociación de Empresas de Acuicultura de España.*
- Arkhipova, I., & Meselson, M. (2005). Deleterious transposable elements and the extinction of asexuals. *Bioessays*, *27*(1), 76-85.
- Baroiller, J. F., D'Cotta, H., & Saillant, E. (2009). Environmental effects on fish sex determination and differentiation. *Sexual development*, *3*(2-3), 118-135.
- Benzekri, H., Armesto, P., Cousin, X., Rovira, M., Crespo, D., Merlo, M. A., ... & Ponce, M. (2014). De novo assembly, characterization and functional annotation of Senegalese sole (*Solea senegalensis*) and common sole (*Solea solea*) transcriptomes: integration in a database and design of a microarray. *BMC genomics*, *15*(1), 952.
- Chalopin, D., Volff, J. N., Galiana, D., Anderson, J. L., & Schartl, M. (2015). Transposable elements and early evolution of sex chromosomes in fish. *Chromosome research*, 23(3), 545-560.
- Charlesworth, D., Charlesworth, B., & Marais, G. (2005). Steps in the evolution of heteromorphic sex chromosomes. *Heredity*, *95*(2), 118.
- Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, *70*(6), 066111.
- Colen, R., Ramalho, A., Rocha, F. and Dinis, M. (2014). *FAO Fisheries & Aquaculture Solea spp.* Fao.org. Disponible en: http://www.fao.org/fishery/culturedspecies/Solea_spp/en
- Dinis, M. T., Ribeiro, L., Soares, F., & Sarasquete, C. (1999). A review on the cultivation potential of *Solea senegalensis* in Spain and in Portugal. *Aquaculture*, *176*(1-2), 27-38.
- Dressman, D., Yan, H., Traverso, G., Kinzler, K. W., & Vogelstein, B. (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences*, *100*(15), 8817-8822.
- Fablet, M., & Vieira, C. (2011). Evolvability, epigenetics and transposable elements. *Biomolecular concepts*, *2*(5), 333-341.
- Fedurco, M., Romieu, A., Williams, S., Lawrence, I., & Turcatti, G. (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic acids research*, *34*(3), e22-e22.

- Fernández, I., Ortiz-Delgado, J. B., Darias, M. J., Hontoria, F., Andree, K. B., Manchado, M., ... & Gisbert, E. (2017). Vitamin a affects flatfish development in a thyroid hormone signaling and metamorphic stage dependent manner. *Frontiers in physiology*, *8*, 458.
- Fernández-Díaz, C., Yýfera, M., Cañavate, J. P., Moyano, F. J., Alarcón, F. J., & Díaz, M. (2001). Growth and physiological changes during metamorphosis of Senegal sole reared in the laboratory. *Journal of Fish Biology*, *58*(4), 1086-1097.
- Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, *9*(5), 397.
- Flutre, T., Duprat, E., Feuillet, C., & Quesneville, H. (2011). Considering transposable element diversification in de novo annotation approaches. *PloS one*, *6*(1), e16526.
- Froese R., Pauly D. (2003). FishBase. http://www.fishbase.org/summary/Solea-senegalensis.html
- García-Cegarra, A., Merlo, M. A., Ponce, M., Portela-Bens, S., Cross, I., Manchado, M., & Rebordinos, L. (2013). A preliminary genetic map in *Solea senegalensis* (Pleuronectiformes, Soleidae) using BAC-FISH and next-generation sequencing. *Cytogenetic and genome research*, *141*(2-3), 227-240.
- Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, *99*(12), 7821-7826.
- Goecks, J., Nekrutenko, A., & Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, *11*(8), R86.
- Guzmán, J. M., Ramos, J., Mylonas, C. C., & Mañanós, E. L. (2011). Comparative effects of human chorionic gonadotropin (hCG) and gonadotropin-releasing hormone agonist (GnRHa) treatments on the stimulation of male Senegalese sole (*Solea senegalensis*) reproduction. *Aquaculture*, *316*(1-4), 121-128.
- Heule, C., Salzburger, W., & Böhne, A. (2014). Genetics of sexual development: an evolutionary playground for fish. *Genetics*, *196*(3), 579-591.
- Hollister, J. D., & Gaut, B. S. (2009). Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome research*.
- Imsland, A.K. 2010. The Flatfishes (Order: Pleuronectiformes). In: François, N. Le, Jobling, M., Carter, C., Blier, P. (Ed), *Finfish aquaculture diversification*. 681 p, CABI, Wallingford, UK.
- International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860.

- Ivancevic, A. M., Kortschak, R. D., Bertozzi, T., & Adelson, D. L. (2018). Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. *Genome biology*, *19*(1), 85.
- Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research*, *110*(1-4), 462-467
- Jurka, J., Kapitonov, V. V., Kohany, O., & Jurka, M. V. (2007). Repetitive sequences in complex genomes: structure and evolution. *Annu. Rev. Genomics Hum. Genet.*, *8*, 241-259.
- Klaren, P. H., Wunderink, Y. S., Yufera, M., Mancera, J. M., & Flik, G. (2008). The thyroid gland and thyroid hormones in Senegalese sole (*Solea senegalensis*) during early development and metamorphosis. *General and Comparative Endocrinology*, *155*(3), 686-694.
- de Koning, A. J., Gu, W., Castoe, T. A., Batzer, M. A., & Pollock, D. D. (2011). Repetitive elements may comprise over two-thirds of the human genome. *PLoS genetics*, *7*(12), e1002384.
- Kuraku, S., Zmasek, C. M., Nishimura, O., & Katoh, K. (2013). aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity. *Nucleic acids research*, *41*(W1), W22-W28.
- van de Lagemaat, L. N., Landry, J. R., Mager, D. L., & Medstrand, P. (2003). Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends in Genetics*, *19*(10), 530-536.
- Lan-Chow-Wing, O., Confente, F., Herrera-Pérez, P., Isorna, E., Chereguini, O., Rendón, M. D. C., ... & Muñoz-Cueto, J. A. (2014). Distinct expression profiles of three melatonin receptors during early development and metamorphosis in the flatfish *Solea senegalensis*. *International journal of molecular sciences*, *15*(11), 20789-20799.
- Laudet, V. (2011). The origins and evolution of vertebrate metamorphosis. *Current Biology*, 21(18), R726-R737.
- Lippman, Z., Gendrel, A. V., Black, M., Vaughn, M. W., Dedhia, N., McCombie, W. R., ... & Carrington, J. C. (2004). Role of transposable elements in heterochromatin and epigenetic control. *Nature*, *430*(6998), 471.
- Manchado, M., Salas-Leiton, E., Infante, C., Ponce, M., Asensio, E., Crespo, A., ... & Cañavate, J. P. (2008). Molecular characterization, gene expression and transcriptional regulation of cytosolic HSP90 genes in the flatfish Senegalese sole (*Solea senegalensis* Kaup). *Gene*, 416(1), 77-84.
- Molina-Luzón, M. J., López, J. R., Robles, F., Navajas-Pérez, R., Ruiz-Rejón, C., De la Herrán, R., & Navas, J. I. (2015). Chromosomal manipulation in Senegalese sole (*Solea*

senegalensis Kaup, 1858): induction of triploidy and gynogenesis. *Journal of applied genetics*, *56*(1), 77-84.

- Novák, P., Neumann, P., & Macas, J. (2010). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC bioinformatics*, *11*(1), 378.
- Novak, P., Neumann, P., Pech, J., Steinhaisl, J., Macas, J. (2013) - RepeatExplorer: A Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next generation sequence reads. *Bioinformatics*
- Okonechnikov, K., Golosova, O., Fursov, M., & Ugene Team. (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*, *28*(8), 1166-1167.
- Oliver, K. R., & Greene, W. K. (2009). Transposable elements: powerful facilitators of evolution. *Bioessays*, *31*(7), 703-714.
- Ono, R., Kobayashi, S., Wagatsuma, H., Aisaka, K., Kohda, T., Kaneko-Ishino, T., & Ishino, F. (2001). A retrotransposon-derived gene, PEG10, is a novel imprinted gene located on human chromosome 7q21. *Genomics*, *73*(2), 232-237.
- Portela-Bens, S., Merlo, M. A., Rodríguez, M. E., Cross, I., Manchado, M., Kosyakova, N., ... & Rebordinos, L. (2017). Integrated gene mapping and synteny studies give insights into the evolution of a sex proto-chromosome in *Solea senegalensis*. *Chromosoma*, *126*(2), 261-277.
- Robledo, D., Hermida, M., Rubiolo, J. A., Fernández, C., Blanco, A., Bouza, C., & Martínez, P. (2017). Integrating genomic resources of flatfish (Pleuronectiformes) to boost aquaculture production. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics*, *21*, 41-55.
- Ruiz-Ruano, F. J., López-León, M. D., Cabrero, J., & Camacho, J. P. M. (2016). High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Scientific reports*, *6*, 28333.
- Serrato-Capuchina, A., & Matute, D. R. (2018). The Role of Transposable Elements in Speciation. *Genes*, *9*(5), 254.
- Shah, A. B., Schielzeth, H., Albersmeier, A., Kalinowski, J., & Hoffman, J. I. (2016). High-throughput sequencing and graph-based cluster analysis facilitate microsatellite development from a highly complex genome. *Ecology and evolution*, *6*(16), 5718-5727.
- Sotero-Caio, C. G., Platt, R. N., Suh, A., & Ray, D. A. (2017). Evolution and diversity of transposable elements in vertebrate genomes. *Genome biology and evolution*, *9*(1), 161-177.
- UniProt Consortium. (2016). UniProt: the universal protein knowledgebase. *Nucleic acids research*, *45*(D1), D158-D169.

- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., ... & Gocayne, J. D. (2001). The sequence of the human genome. *science*, *291*(5507), 1304-1351.
- Vinogradov, A. E. (2004). Evolution of genome size: multilevel selection, mutation bias or dynamical chaos?. *Current opinion in genetics & development*, *14*(6), 620-626.
- Volff, J. N., Körting, C., & Schartl, M. (2001). Ty3/Gypsy retrotransposon fossils in mammalian genomes: did they evolve into new cellular functions?. *Molecular biology and evolution*, *18*(2), 266-270.

ANEXO

AI. GRÁFICAS DE CALIDAD DE LOS DIFERENTES PAQUETES DE DATOS

AI.I. Paquete de datos S0

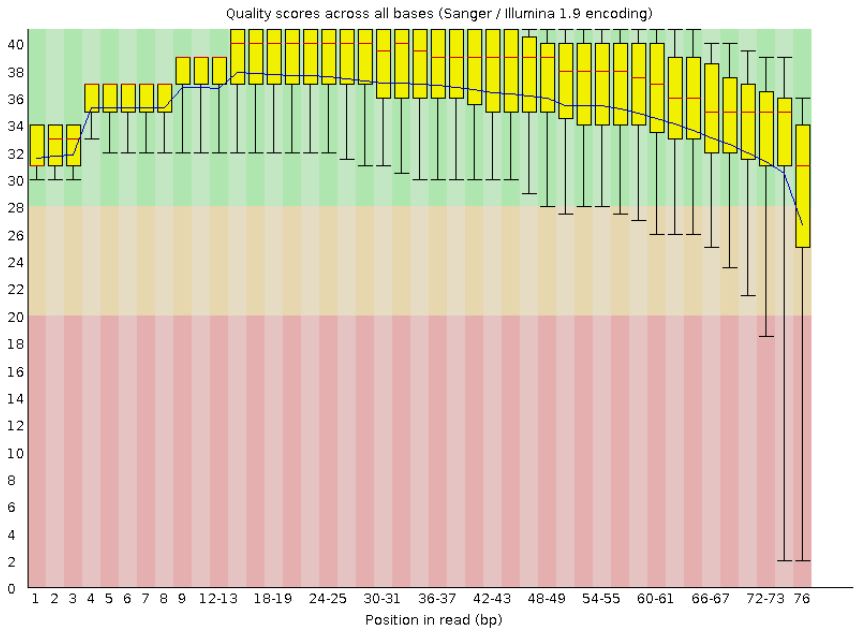


Figura I.I.I. Valores de calidad de las secuencias por base de las lecturas del paquete S0.

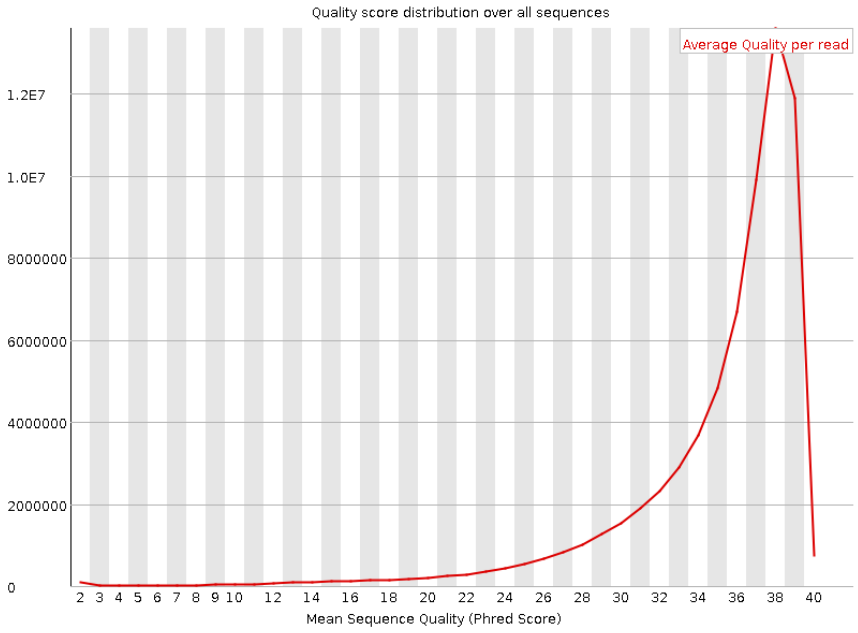


Figura I.I.II. Valores de calidad por secuencias de las lecturas del paquete S0.

AI.II. Paquete de datos S4

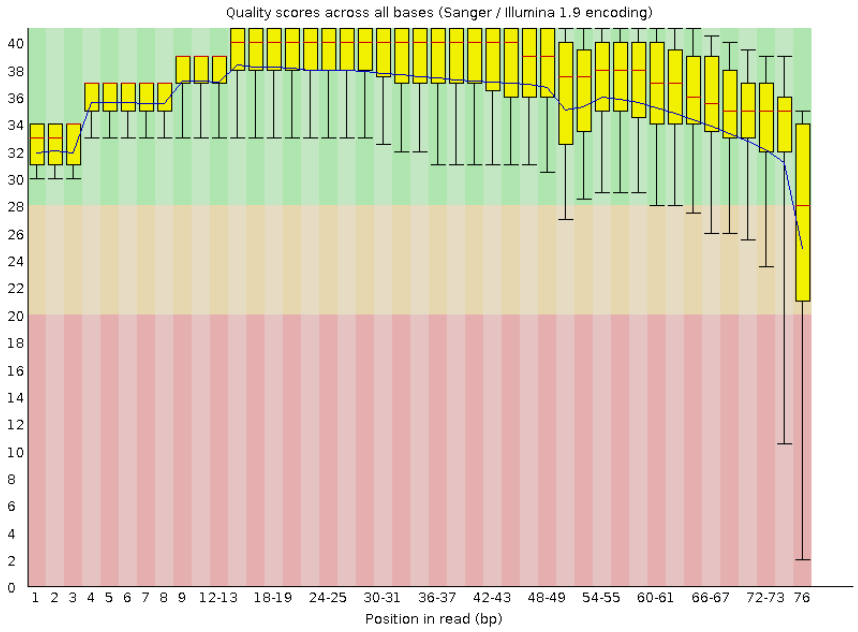


Figura I.II.I. Valores de calidad de las secuencias por base de las lecturas del paquete S4.

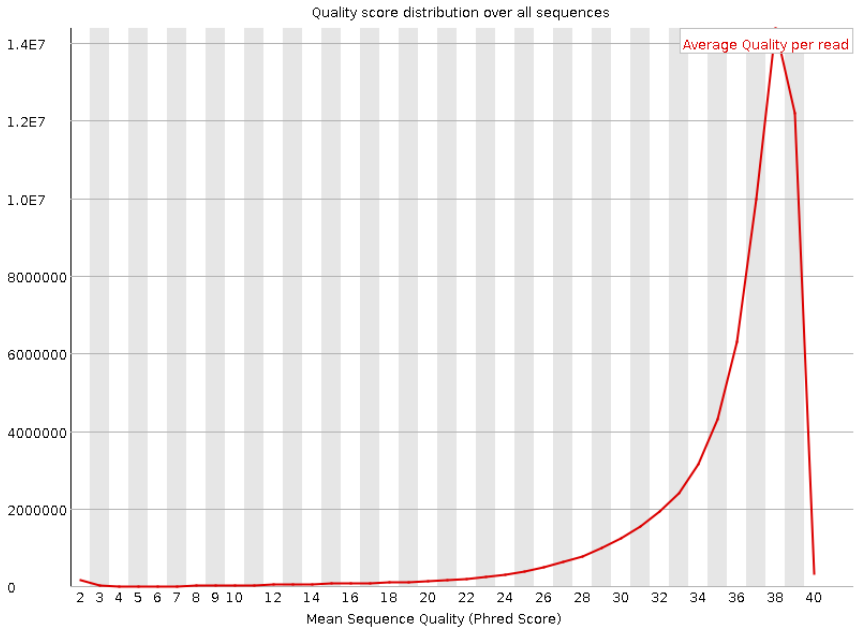


Figura I.II.II. Valores de calidad por secuencias de las lecturas del paquete S4.

AI.III. Paquete de datos gónadas

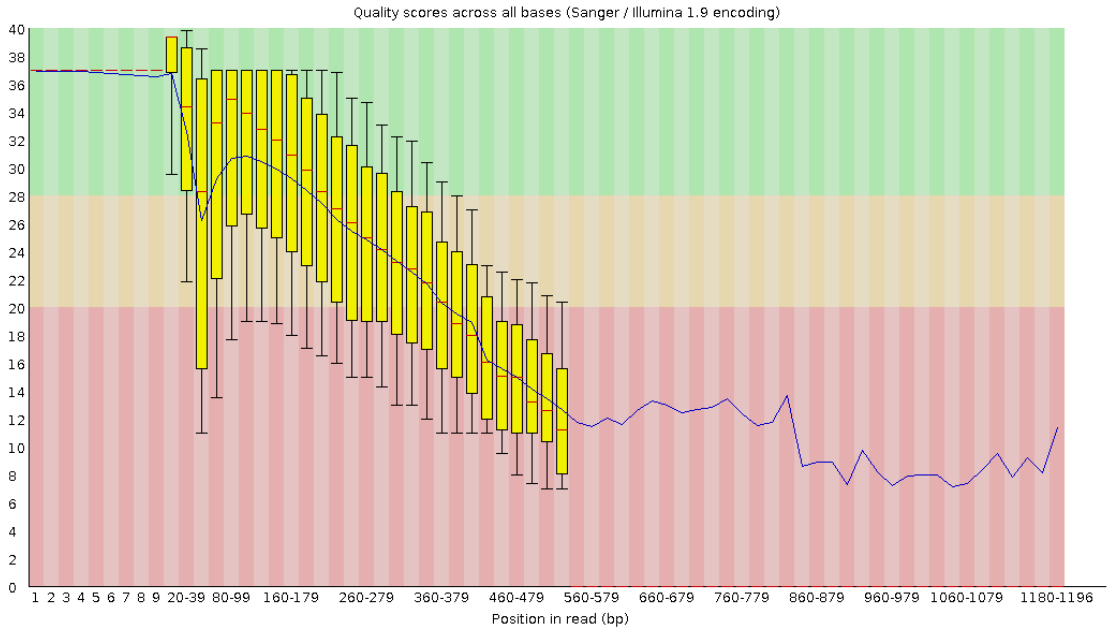


Figura I.III.I. Valores de calidad de las secuencias por base de las lecturas del paquete gónadas.

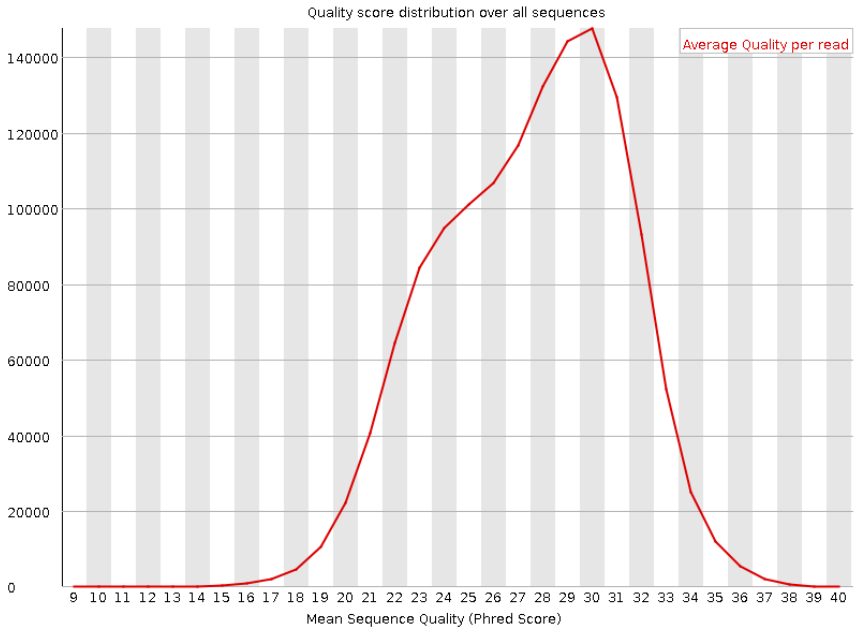


Figura I.III.II. Valores de calidad por secuencias de las lecturas del paquete gónadas.

AI.IV. Paquete de datos sistema inmune

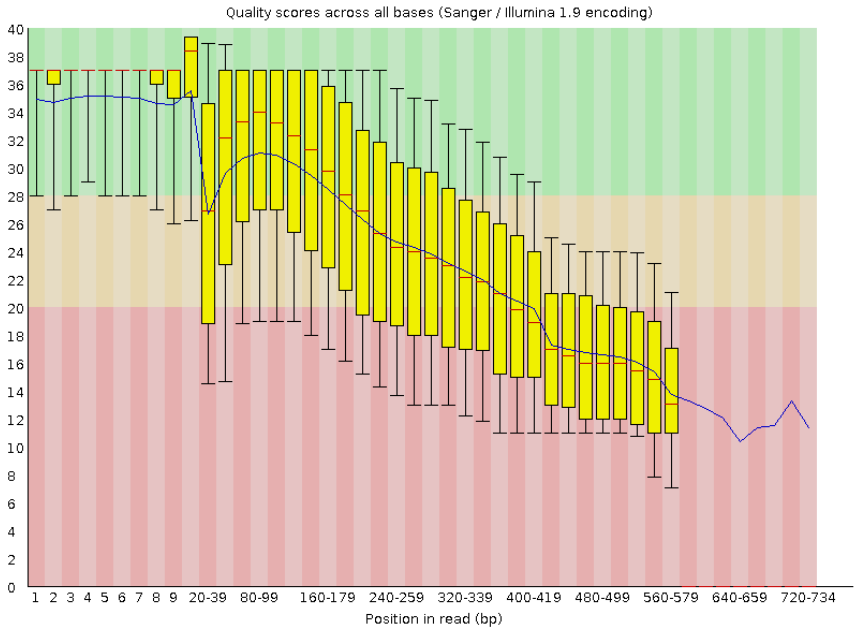


Figura I.IV.I. Valores de calidad de las secuencias por base de las lecturas del paquete sistema inmune.

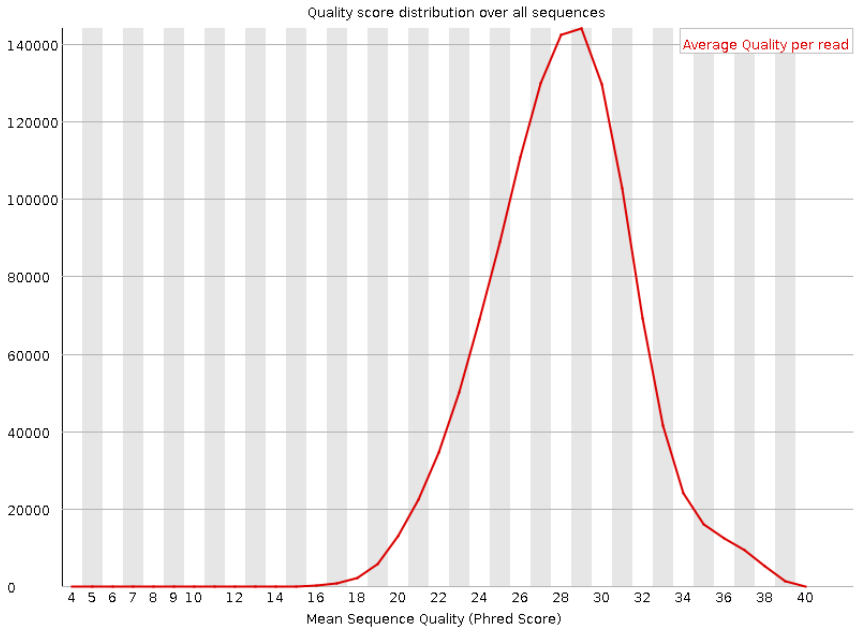


Figura I.IV.II. Valores de calidad por secuencias de las lecturas del paquete sistema inmune.

AII. SECUENCIAS CONSENSO DE LOS CLUSTERS DE NIVEL 1

AII.I. Secuencias consenso de los clusters de nivel 1 seleccionados en el paquete S0

>Sse_LarvaS0_ClS0_1:

AAATGTGACAAAAAAGTCATAGTTTAGTATGTCGTCCA

>Sse_LarvaS0_ClS0_2:

TTGTAGTTGTTGTGGCTGGTG

>Sse_LarvaS0_ClS0_3:

TGAGGCAAACAACCCCAGCGGCTTTACCCTGGATGANGTCATCCAGACTGGTGTTGACAACCCAGGTCACCCCTTCATCATGACTGTG
GGCTGCGTCGCTGGGGACGAGGAGNCNTATGAGGTCTTCAAGGAGCTGCTGGACCCCTCATCTCCGACCGTCATGGTGGATACAACCA
CCGACAAGCACAGACCGACCTGAACTTCGAGACCTGAAGGGTGGTGATGACCTGGACCCCAACTAGTTGTCCAGCCGGTCCGTACTGG
CCGCAGCATCAAGGGATTCACCCTGCCCCCCACAACAGCCGTGGCGAGCGCAGAGCCATTGAGAAGCTGTCCATTGAGGCTCTGAACA
GCCTGGATGGCGAGTTCAAGGGCAAGTACTACCCCCTGATATGACTGATGCGAGCAGGAGCAGCTGATCTGACACTTCCTGTTTGACA
AGCCCGTTCCCCCCTGCTGACCTGCGCGGATGGCCCGGACTGGCCCGATG

>Sse_LarvaS0_ClS0_4:

CGTATCAGAACTCACCTGTAGTTTNAGTTAGCTGNANAANGAGAGTTCTNTTTCTNACTTNTCNTCTAATCCTGAAGTNANAGTTCNG
TTCTGTTGTTGNCGNACTCACTTTATNAGNTGATGATTTCCGGTTCAGNTNGANAGGATTGGNNGTTTTGAGGACACCACCNTTTCCA
TCGACNATCTCCTCTATGACCACCATGATCTTTGTGGTGGTCCTGGAGGAAGAGGATNTNCCATTACTNCATCCTGNTANCCTCTCCG
TCCAGCAGCCTCCTGTACTCTGCGATCTCCATCTCCAGTCTGGTCTTGATGTCCAGCAGCTATCTGGTACTCNTGGCTCTGGCTCTCC
ANGTCAGCCCTCAGCTGCNCCAGCTGCTCCTCCAGANNNATNACCTGGATCTGGAGGNTCNTGAGCTGCATGGCGTAGCGGTTCTGTG
TCTCNNNCANTNTNTNTTCCAGAGNCGCTTTCATGCTAGTGGTGAGTCGATCTCCAGCTAGCATCTATCGCTGAGTGCGAGTC

>Sse_LarvaS0_ClS0_5:

CCCTGCGTCAGAGAGCAGAAGGGACTATAACATTGGCTGCATTTTATTCAAGAAAAGTCAACGAAATAAACAAGATGCCTGAAGAAAT
GCGCCAAGAGGAGGAGACTGAGACCTTTGCCTTCCAGGCAGAGATCGCTCAGTTGATGTCCCTGATCATCAACACTTTCTATTCCAAC
AAAGAGATCTTCCTCAGAGAGTTGATCTCCAATGCCTCTGATGCTTTGGACAAAATCCGCTATGAAAGCCTGACTGATCCAACCAAGC
TGGACAGTGGCAAGGATCTGAAAATTGATATCATCCCCAACAAAGCTGACCGAACCCTGACCCTCATCGATACTGGAATTGGTATGAC
CAAAGCCGATCTGATCAACAACCTGGGTACCATTGCCAAGTCTGGCACCAAGGCCTTCATGGAGGCCCTGCAGGCTGGAGCTGACATC
TCCATGATTGGTCAGTTTGGTGTGGGTTTCTACTCGGCTTACCTTGTTGCTGAGAAGGTAGTTGTCACCACCAAACACAATGATGATG
AGCAGTATGCCTGGGAGTCTTCTGCTGGAGGTTCCTTCACTGTCAAGGTCGACACAGGTGAGCCCCTTGGCCGTGGAACAAAGATTGT
CCTGCACCTGAAGGAGGACCAGACTGAGTACACTGAGGACAAGAGGGTCAAGGAGATTGTCAAGAAGCACTCACAGTTCATTGGCTAC
CCCATCACCCTGTTTGTGGAGAAGGAGCGCGACAAGGAGATCAGTGACGACGAGGCAGAGGAGGAAAAGGCTGAGAAGGAGGAGAAAG
AGGAAGAGGAGGCAGAGGACAAGCCAAAGATTGAGGATGTGGGCTCAGACGATGAGGAAGACTCCAAAGACAAAGACAAGAAGAAGAA
AAAGAAGATCAAGGAGAAGTACATCGACCAGGAGGAGCTGAACAAGACCAAACCCATCTGGACCAGAAACCCTGATGACATCACAAAT
GAGGAGTATGGAGAGTTCTACAAGAGTCTGACCAATGACTGGGAGGATCACCTGGCTGTCAAGCACTTCTCAGTGGAGGGTCAGCTGG
AGTTCAGGGCTCTTCTCTTCATCCCCCGTCGTGCTCCTTTTGATCTCTTTGAGAACAAGAAAAAGAAGAATAACATCAAGCTGTACGT
CAGGAGAGTTTTCATCATGGACAACTGTGAAGAACTTATCCCAGAATACCTGAACTTTGTCCGTGGTGTGGTCGACTCAGAAGATCTG
CCCCTCAACATCTCCAGAGAAATGCTGCAGCAGAGCAAGATCCTCAAAGTCATTCGCAAGAACATCGTCAAGAAGTGTCTGGAGCTCT
TTGCTGAGCTGGCTGAGGCCAAGGAGAACTACAAGAGTTTCTACGAAGCCTTCTCCAAGAACATTAAGCTCGGTATCCATGAGGACTC

GCAAAACCGCAAGAAGCTTTCTGAGCTGCTGCGTTACCACAGCTCCCAGTCTGGAGACGAGATGTCCTCCCTCACAGAGTACATTTCT
CGTATGAAAGACAACCAGAAGGCCATCTACTACATCACTGGAGAGAGCAAGGATCAGGTGGCCAACTCTGCCTTTGTTGAGCGTGTCC
GCAAGCGTGGCTTTGAGGTCCTGTACATGACCGAGCCAATCGATGAGTACTGTGTCCAGCAGTTGAAGGAGTTTGATGGTAAGAACCT
GGTCTCTGTCACCAAGGAGGGTCTGGAGCTTCCTGAGGATGAGGAGGAGAAGAAAAAGATGGAGGAAGACAAGGCCAAGTTTGAGAAT
CTCTGCAAACTCATGAAGGAGATCCTTGACAAGAAAGTGGAGAAGGTGACCGTGTCCAACAGACTGGTGTCTTCACCCTGCTGCATTG
TGACAAGTACTTATGGCTGGACGGCCAACATGGAGAGGATCATGAAGGCCCAGGCTCTCAGGGACAACTCCACCATGGGCTACATGAT
GGCCAAGAAGCACCTTGAGATCAATCCTGACCACCCCATCGTGGAAACTCTCAGGCAGAAGGCCGAAGCCGACAAAAACGACAAGGCT
GTGAAGGACCTTGTCATCTTGCTGTTCGAAACTGCCCTGCTGTCCTCAGGCTTCTCCCTGGATGACCCACAGACTCACTCCAACCGCA
TCTACAGAATGATCAAACTCGGACTGGGTATCGATGACGATGAAGTTCCCACAGAGGAGACCACCGCCACATCTGTCCCAGATGAGAT
TCCTCCGCTCGAGGGTGAAGGCGAGGATGACGCTTCACGCATGGAAGAAGTCGATTAAACCATCTTTCTCCTCTTCCCAGATTTTTAA
CACTTTAACCTCACTTTTCAATTGTTCATCCCTAAACCTGCAGTAATTGCAAAACAAATAGTCATTCATGTTGTGCGGTTGGCCAGTG
TTGCTCCTGTGTACAGAGCAGTTACTCTGCAACACCATTTTAAGAAAAGCAATTTTGGTTTTTGCTCTACAAGTTCATGGTGACAGCA
CATTTGTTTATCAAGTACCCTGTTGCACTGAGTTTTAAATGTTGGAAAATGTGTGAACATGGGAATGCTACATTCCAGTATAAGGGTC
AGGAGGGTTTGTGGAGGTTCTGCTCATGTGCAACACTGCACGCTGCATGGAGAGAGGACTGTATGATTCCTTTGCCTGAGTCCAGGCT
TGTCTGTATTCCAAGTCTTTGTTTTGCAAAAAAT

>Sse_LarvaS0_ClS0_6:

AACAAAACAGGGTTAGTTTGAGGTAAGAACAAAAACTATGTGAAACACGTTTGACTGTGGTGTGTGCGACATTCTGCAGAGTCCTGTG
AATAAAACCTAGTGATC

AII.II. Secuencias consenso de los clusters de nivel 1 seleccionados en el paquete S4

>Sse_LarvaS4_ClS4_1:

ACATACTAAACTATGACTTTTTTGTCAGATTTTGGACG

>Sse_LarvaS4_ClS4_2:

GCTCCTCAGGATTCTCTGGTTCTTCTG

>Sse_LarvaS4_ClS4_3:

CACAGGTCCTGTCGTTACAACTGTGGTTATGACCTGGGAGTCTGCTCCTGCCGCAGCTCGTGTAGATACAATGGTACCTGTTGCTATG
ACTACAACAACTACTGCCTTGGAACGACAGCACCACCAGAAACAACAGGA

>Sse_LarvaS4_ClS4_4:

TACTACAACCACAACTCCAACTACTACAACCACAACTCCCAC

>Sse_LarvaS4_ClS4_5:

TCCAGTTCCTCCTTCCAGTTCAACCTCTGGACCTTCAATCTTTCCTTTGGGCAAAGAGATGTCAATATTTGGCATTTTCATGTGAGGC
ATTTTGAATTTTCC

>Sse_LarvaS4_ClS4_6:

GTCAAGAAACCATCTAAAGAAGAGAAAGAGGTGAAGCCTACCAAAGAAAAGAAAGAA

>Sse_LarvaS4_ClS4_7:

CCGAGTGTTGACGTCAAAGCTCCAGAGATTGACATCGAAGCTCCTGATGTAAAACTCCATGGACCAAATATCAAATTACCATCAATTT
CAGCGCCCAAGCTCCCAGACTGGGATCTTAAACTGAAAGGGCCCAAAGTAAAGGGAGATGTTGATGTCTCAGTTCCAAAGATTGAGGG
TGATATAAAAGGACCCAAACTTGATATTGAAGGACCAGATGTTGACCTTGATGGTAAAACAGGAGGATTTAAAATGCCTAAATTCAAA
ATGCCATCCTTTGGATTTAAAGGCTCACATGGTGAAGGGCCAGAGGTTGATGTTAGTCTCCCGGAGGCTGATATTGATATCAGAGCTC
CAGATATTGATATCAAAGGACCAGAGGTTGACCTGGAAAGTCCCAGTGGAAAGATCAAGGGATCAAAATTCAAAATGCCAAACATCAA
AGGACCTCAAATCTCTATGCCTGATGTGGATTTTAATTTGAAAGGTCCAAACTGGAAAGGCGGTGTGGATGTTTCAGGTCCAAAGATT
AAAGGAGACATAGGAAAACCTGACATTGATTTCAAAGGTCCAGGGATTGATATTGAAGGACCAAAGGCTGGATTTGAAATGCCTAAAA
TCAAAATGCCAACTTTCAAAGGTGCTAAAATGGAGGGCCCAGATATTGATGTGAACCTCCCTAAGGCTGACTTTGATGTGAACTTA

>Sse_LarvaS4_ClS4_8:

CTTTGGCCTGAGGGTGCGTTTCGATGGTAACCACCATGCTGACGTCACCTTGCCGACCTCCTACAGTGGCCTGCTCTGTGGCATGTGT
GGAAATTTCAATAACAATCCAAGAGACGACAACCTGAAACCAGACCAAACACCAGCCGCTAACACCAATGAGTTGGGAGACAGCTGGC
AGGTTCCTGACCCGCGGCCTGACTGCACCAACGGTGGAGGACATGAAGAGTGTGACAAGAATGTGGAGGAGGAGGCCCAGAAACCAAC
CAGCTGTGGCATGATCACCGATCCTAACGGTATCTTCAAGCCTTGCCACTCTGTCGTGCCCCCGAACCAATACTTTGAAAACTGTGTG
TACGACGTGTGTGCAAATGGAGGTCAGACTGAGGCTCTGTGCCAGGCCATAGAGAGCTACGCTGATTTGTGTGCTGCAGCAGGAGTCC
CCATCGCATGGAGGAAAAACAACACCTTCTGTCCTATCAAGTGTCCCTCAGGCAGTCAGTACAATCCATGTACCTCTGCGTGTCCTCA
GCCCAGTTGCCAGGACCCTGCAGGCTCCGGTGGCTCCTGTAATCAGCCCTGTGTGGAGGGATGTGTCTGTAATCCTGGGCTCATCCTC
AGTGGAGACAAATGTGTCCCGCTCAGTGAGTGTGGATGCACTGATGAAGGTGGAAATTACAGGCCGACAGGAGACACTTGGTTCTCAG
AGAAGGACTGTTCAGAGCGCTGTAAGTGTAACGGCAACCACAACATCACCTGCGAGCCATGGCAGTGCAGCCCTACACAGGAGTGTAA
GGTGGTGGAGGGAGTACTAGGCTGTTACTCTAGAGGAAATGGAATCTGCTCAGTATCTGGTGATCCTCACTACAACACCTTTGACAAA
GTAACCCACCACTACATGGGGTCTTGCTCCTACACCTTGACCAAACCCTGCAACGTCTCCACCGACTTGCCGTACTTCACCGTGGACA
CCCAGAACGAGCACAGAGGAAGTAACAAGAGGGTTTCCTATGTCAGAGCTGTAGTGATCAATGTGAGCGGTGTGACTGTCATCCTTGG
CAAGGGACGCAAAGTCCAGGTCAATGGGACAGCAGTCGTCCCACCTTTGAATCCTGCCAAAGGAGTCAAGATCTACTTAAGTGGAAAG
TTTGTCGTCCTGGAGACAGA

AII.III. Secuencias consenso de los clusters de nivel 1 seleccionados en el paquete gónadas

>Sse_Gonadas_ClGo_1:

TGATGAGCATGGGGAAGGCTGCTTTCCAACCATCAACCCTTTCCGCTACACAGAGACTCTGCACTTCAGCCATGGGGACAACCAGTCA
TCCACTTCATGTTCAATGCCTTCCATGCAGAGTCTAAAAAGCCTCTGCACAGGGAGTGTGGCTTCATTCGAATGCAGCCAGAAACCAA
CAAAGTGATGTTTATCTTGCACAAAACACACTCAGCCACTTAGTAGTGTGGCGTTCTGAGACACAGAATCACTACTTTGACTTTA

>Sse_Gonadas_ClGo_2:

AGAGTACGGGGGATGACGGCAGCATGCGCATTTCACAGCGAATGTTACCGCCCACCAGCGCCTCAGAGTACGGGGGCCCTTCTTCCGG
TCCTCCTCAGACGTGTTTCGGTCCACGTCCTCCGTCTCTTCTTCATCATGGGTTTCGGTGACTTGAAATCATCTTCAGGTCTCGCGCT
GCTGAACGACTTCCTGTCTGCTCGCAGTTACGTTGAAGGGTATGTCCCCTCTCAGGCCGACGTGGCCGTCTTTGACGCCATCTCGACT
CCGCCCTCAGCCGACCTTTGCCACGCCCTCCGCTGGTATAACCACATCAAGTCGTACCAGGACCAGAGAAACAGTCTTCCAGGTGTGA
AGAAACCACTGGGACACTACGGTCCTGCAGGCGTAGCCGACACCACCACAGTCTCCGCCCCCACGTCAAAGGACGATGATGACGATGA
CGACATTGACCTTTTCGGTTCTGATGAAGAGGAAGACGCAGAGGCGGAGAAGCTGAAGGAGCAGCGTCTCGCGGAGTACGCCGCCAAG
AAGTCAAAGAAGCCCACCCTCATCGCCAAATCCTCCATCTTGTTGGACGTGAAGCCATGGGACGACGAGACGGACATGTCCAAACTGG
AGGAGTGTGTCCGCAGCGTCCAGATGGACGGACTCGTCTGGGGACAGGCCAAACTGGTGCCAGTGGGTTACGGCATCAAGAAGCTTCA
GATTGGCTGTGTGGTGGAAGACGACAAGGTGGGCACTGACATCCTGGAGGAAAAGATCACGGCCTTCGAGGACTTTGTTCAGTCCATG
GACGTCGCTGCCTTCAACAGATCTGAAACTAACAAATAAAAGCACTTC

>Sse_Gonadas_ClGo_3:

AGAGTACGGGGGAGTCATGATGGCTTCCTTTTGGCAAGGTGTGCTCCTCTTGAGTCTCATCACTATTTCAGTTTATGCAGACATGAAG
CTGGACTGTAGGCCTGATTTTGTGACACTAGTGTGGACAGAGAGCGGAAGCCAGGTTGACTTGTCCCTCTTTCGTCTTGGTAGCTGTT

TCCCCACCAGCATCACGGCCAGGGAGGTTGTTTTTAGCGTGGACTTCAGTGACTGTTATTTCAGCAGACTGGTCACTGGGGATCATCT
GATATACGCCAACAACCTGACCTACATTTCTACCGCCGAGTCTCGTATACACTCCTTCAGTCACCCGGTTGTTTGTACATATGATAGG
CCTAAAGACTGGTACCCACTCATTTACCAACCCGTGTTTAACACACATGGTCAAGGAGATCTAATTTTCCATGCTGGACTTATGAACG
ATGACTTCTCAGGACCTGCTGTGTTGACAAGTTTCCCTCTGGGCTCCTTCATCCCAATCATGGCCAGTGTGGAGCAAGCCACCCATCA
GCCCTTGCTGCTGCTTCTTGAGGAATGTGTTGCTGCTGCTACACCAAAGCTGCACCCTGAAAGCAACTTGTATACAATAATCACAAAT
AAGGGATGTCTTGTGGACAGTAAGATCTCCCGCTCAAAATTTGAACCAAGGCAGAAATCATCCGAGATCAAATTATCCCTTCAAGCCT
TCAGGTTTGGTCTTGGAGAACAGGTGTATCTCCATTGCAAACTTGTGGCTTGGGATCCCATGGGTCTGGACAACACAAAGAAGGCCTG
TCACTATGTCGGAGATCAGGGCTGGGAGCTTCTGGATAACCCGGCATACAGTAATCTGTGTGATTGCTGTGACTCAAGCTGCGGGTCT
AGGAAAATAAGGAGCACAGGGTTGGGTAAACTTGGTATTGCACATGGAGCAGTCCTTGGGCCTTTCACCATCACAGAGTCCTGAAGCA
ACTAAGTGACACTGCTTGGTTTTTGACATTGCACAAGATTTATTTGTGCTTTGAACTTTATTTTCCCCTCTGCAGAACTGAATCTATC
GTATTGCCAACCTGCTTCAGTCTGAGTTTTTAGTGTTGTCAGGCTTCTAAAAGCTTAAATGTCTGCCTTTCAGATGTGTGGCAATTAT

>Sse_Gonadas_ClGo_4:

AGCACTCAGACTGGCAGGTATTTCAGAGAGAGGAGCAGTTATCCCATGAACCTTCGAGCTCGAACAGGGTAGGTTTCCTGACAGAGAA
GCTTGTTTTCGTACACCTAGGCAACCGACTCATTTAGCAGAGGTGAATGTGTTCCTGTAGAGGTGGTCCATGCACTTGTTCTCCAGAC
AAAGTGGGTTGTCCTCTGTTGACCTTGTAGCTGCCTGCTCATGTATGAAGACATATGATGGTGGATACATGGTATGGAAAACCCCTGA
GTGCTGCACCCCTGGTGTCAGATGTCATGACACACAGGTCAACATTGGTGTCAATGGTGAACTTGTGGAGCAGCCAGTGGCAGAGGAG
CGAGGCTACGTTATTGAGAAGCACAATGCTACAGTTGAGATCCACATCCCTTATAATGCTGAAGGAGGACACAAGAAGAGCCTTGTGA
TTGGCGACCTGTACGAGTACTACATCTTTGAACTTTATCTGGAGCAAATCTCGTGGATGAGGATCACGTTGACACCAGGCTACGCTTT
CACAGGACACTTGCCACTCCTCTGCTCCACCCTTTTTACAAACCAACCTGAGAATTTCACCGTCTACCTGGGAGATGTCCCCGAGGAT
GTTGGCTGACTGCTGTGGGTTGAATGGACACACATTCACGCTCCGTTTACAAATGCAAGCGGCTACAACATCCTAAGTTGTTCACCCA
ACAACACCCATGGCTACACTCTGACAGTGCCTTGATGACCCTGTTGTCGTACAGCAGTTCTCCAAAGAAATGCAGCATGAAGCACTGT
GGCGTCAACTACACACTGAGGTTCTGCCTGATAATGGGCCTTACTACCACCAGAGTCGTCTGGCACTAACTATGCCTCTCCTCCAGGT
TTGAGCCTGTCAGAGTTGGCTCGCTTCAACTCACGGCCTTGACTACCTGTGGTACATGTATCGGTTCAGACCGGCTGACACCAGCTGG
CAGCTCGCACAACTACACATGACAAGACAGCCAGAGTTTGCTCGTCCCTCTCACTCATGGATAGAATACAAGACATTTCTTTGAAGGA
TTTCTTGGCACTTTTGAGATCCTTGTGAGGGATCGTGAAACATCAGAGGTCCGGAGATCAACAGTCAAGACTTGTCCTTTCACTCCTA
CAGAGTTCATCATGTGTTCAACTGATGGCAGGATGACTGTTGTGGCTGACTTATCTTATCTCCAACACGGAGGGATCCCAGCCAGAAG
CAACCTCATGACATATACTGTGGACCCAAAGAAGAGAACACCAGGCTCTTTCACTTCCCTTAATGCTGTGGATCCACAGTCAAGACTT
GGCAAGGACTCTGTGACCTATGAAAATGAGATCTTCTTCAGCAAGAAGTTTCAATCGGATTCCAGCAATGTTGTAGCAGGGGTNAAAC
TNCAGTGTACNTATCTTCTGTCTGGGCTCCATCGGCTCTTCTCGGTGTACAAGTTTGAGTCTGACACAGTTGGCCTTGGACACGTTGT
ACATTCTGCACACTCACTGAAGTAGAGCTACAGGAACCCATGGCGTTCACTGCAGTGCTTCAAACAACGCTGCCTCCTACCAGACGAC
AAGCATTCACCTGCTTTCCGCTCCTGTCACCAATCAAGTTCAATAAATTAGAAGGACTGAAGATCTNCAAAACAAAGTGAACTGTTAT
GTATA

>Sse_Gonadas_ClGo_5:

AGAGTACGGGGGAATTCCTTCGTCGGGCTACTTGAATTGGGAGAGAGCAGAGTACGGGGGGGTAAATGGACCGAGGCTAACAGCTGCG
CCTTTTTAGCGTCTTTTGGAAACGTGTACGAGTATGGCGACTTAAGTTCACATTTTCCGCAAATTGTACGTGTCGAACCGCCGTTTGT
GCGTGTTTCAGTTTGTGTTACGGCTGCTCCCCGTGCCGTTGTTGTCGCCAACAGTTTAGTTTTTTGACTGTAAACGTTTCTGTAACTT
TCGCCTGCAAGGTAGCAGCTCATGCTAGCCTCGGTTTAACCTCTAACTGTCCACTCGAGTGTCATTTTAACTTTACATTAACGAGCAA
AACAGCACACGGCAGCCGCAGGAAAGGTAGTGCAAGGACGGTGTGGCTATGTCAGGGGTAAACAGAGGACAAGCAAACGCGGTTAGCG
AGCAACACAACCAGGAGAATATGCTGTCAAGACTGAGGGGCTCGCTCAAGTCCCGGACTGCCGCTGTAAGCGAAGATCAGGAAAACCT
TCCACCGAAACAAGCATCCAACAGAGCCGTGCTTGCCCCCTTGCAGATCAACAACCAGCGGGGGAAAACTCAGAACCAGCGCGGCACA
AAGCAGGAGTCGTCACAGTCCTTGTCTTGCAAAAATGAAGATTCCAGCAAAAGCTGCCATGAGAAGCCCTCTACCAAGCAGTCTTTCC
AGATCCATGTGGACGAGCCTGATGGTGCCTGCATCAAGAAGCCACTGCAGGTGGTTGAGGCTGTTAAGGCAGAGACTGAGGAATCTCC
ACTGCCCATCGACGATGCCGTGGCACGGCTCAGGCAACCCCTCTCCACCATTGATATTCCTTCGGCAATGGACGTCAGCTTTGACTCA
CCCATGGACATGTCGGTAATTGAGGGGGAGGAGAAACCTGTCAATGTAAATGCAGTCCCAGAATATGCTGCTGAAATCCACACGTACC
TGAGGGAAATGGAGGTAAAAACCAGGCCTAAAGCAGGCTACATGAGAAAACAGCCAGACATCACAAACAGCATGAGGGCCATCCTGGT
GGACTGGCTGGTCGAGGTTGGAGAAGAGTACAAGCTACAGAATGAGACACTTTATCTTGCTGTAAACTACATCGACCGCTTCCTCTCG
TCAATGTCTGTCCTGAGGGGGAAGCTTCAGCTCGTCGGGACTGCTGCTATGCTGCTGGCTTCGAAATTTGAGGAGATCTACCCTCCAG
AGGTGGCAGAGTTTGTTTACATCACAGATGACACCTACACCAAGAAGCAAGTGTTAAGAATGGAGCATCTGGTGCTTAAAGTGCTCTC

CTTTGATCTGGCAGCACCAACTATCAACCAGTTTCTCACTCAGTACTTTGTCAACCAGTCTGTTGGCAAACAGGTGGAAAGTTTGGCC
ATGTATCTTGGGGAGCTCAGTCTGGTTGACTCAGATCTTTTCCTGAAGTACCTACCATCACAGACAGCTGCTGCCGCCTGGGTCCTGG
CCAACCACACAGTGACTGGTGGTTCATGGTCCAAGTCCTTGATGGAGATGAGTGGCTACTCCCTGGATGATCTGATGCCATGTGTTGA
GGATCTGCACCAAACTTACCTCAATGCTTCTCAGCATGCACAGCAATCTGTCCAGGAAAAGTACAAAGGCCCAAAATACCACAGTGTT
TCTACCATCAAAGTGCCAACTAAATTACAGCTGAACTAACTCCTCCCCCTTGTCCATATAGAGTTGCACCTCCTGCCAGTCTTTATTT
TCTATTTTGTTTTTAATCATGTCTACCTGTTTGAAATAACACAGCCTGCTCTTCTTGGAGTAGCGCTAGTCATTTAGAGTTTTTAATG
GTATTTTTATACTCGATCCCCAAACGTTGAAGCCAGATTTCTAGTTGTGTTGGCCAATCAAGCAGTTTTAATGTTGACATTTTAAATT
GAAGGCAATGTACAAGTCTGGGATCACACTGTGTAAACCCCAGTGCTTTGCTCTGCAGTGTCACTGCCACGACATGGTCTCAATCTAC
TCCAACCCAGAAATGTGTCATTTATGCTTCACATTTTCTGGGTTGTGTTATGAGCAGCTGTTGCTGGTTACACTATATCCACGAGGAG
ATGCGACACCGCTTATGTACTGTTAAGTGGTGTAACCTTCAAAGTGTTTGGCCTCATGAGATGCAGATTAAAGTGCATTTTTGTCTTT
CACAGAGTCAACTGTCTCATTACATCTGTCTTGCAATTTGCTTTTCAAGGCTTCTTCTGCACTGTAAGATATGGATCTTGTA

>Sse_Gonadas_ClGo_6:

TTTTTAAACTATCAGCATTTTATTTTCACAATTATACATGTTTTAATTACAATTTTACAAAACAAAGTGTGTACTGCTGCAAAATTCA
GGACATTCAAGGATTCTGCAGATCAAAAGATAACTTCATTTCAGTGCTTCAAGGCATCGTCAGTGTACAGTTTCTGTACAAAAGGATG
TTCAGCATTTGATTTCTCAGTTTAAAAATGAAAGCATGGTGTTTACAACATATGGATTTGGATCCTTTTCTGTCCCCGTGGTCAACTT
CTCAACTCTTCAGATGCCGTTGTACTCACGAGGGCAGTAGCTGTAAATCAGCGGCTCTGCTGGGATTCCACGCTCCACCATGCGACAG
CGGTGCATGGACAGTCTGTTAAAGGCATCATTGAGATCCTTCACATCCATGGCGCTCCACAGCCGCTCGCCCACAGCACTGGCACGAG
GCCAGAGTCGAGGTGTCAGGTTGGTGGCATCCACGTATTCTCCCCACAAGCAAGCTTCTCCACCAATGACTAGTTTCTTCTGCTCCTC
AGTTCCGTTAAAGTCGAGTGGTTCAGCCTTGTAGTAGCGGCCCCAGTCCTGTCCATAGCTAATTAAATCTAGGTACCACGGGGCGGAG
AGGACGGTCTTGTATCCTGCAGCTGTCACCTTGCCCATCTCCTCCTTAGACCCACTGCCGATCCACACATGCACTACTGTATCTGGCT
TTAGCTTAACACCATTGTCAAAAACCTCCTGCCAGATCATGTAGCCCTTATTAGTGGCAGCGACAATGTCCAAGAGTTTTTGGATGTA
GAAGGATTCCAATTTTTTGTAGTCATCTCCAAAGCCCTGCTGAACCATGAACGTCTTAATGTCTGGGTTGGACCTCCAGCAGCTGAAG
TCCACCTCATCACCTCCCAGGTGAATATAAGCGTCAGGGAACACAGAACTGATCTCCTGAAAGAACTGCTTCATAAAGTCATACGTGC
TATTAAGCGTGGGATTCACTGGTCCAAAGGTACCGGAAGGTCCAGCGCCAGTATAGCAGGGGGTGAGCAGATTTGGCTGACCTTTGCC
CCAGGATTGTGTGTGTCCGGGAGTGTCGAACTCTGAGATGACACGAATGCCTCGGAGACGGGCAAACTCCACCACCATCTTCACATCA
GCAGGCGTGTACACATGTGTGTATGGGTGGTAAGCTCCCTTCTGGCTCAGCTGTGGGAATGTTCGGCTCAGGTAAGGGAAGGATTGAT
CGTCAACAATGTGCCAGTGGAAAACGTTGAATTTGTTCATTGCCATAGTTTCCAGAGTGGTCAAGATGACTTTGATGGGCAGGAAATG
GCGGGAACTGTCCAATAAGATGCCTCTGTGTGCAAATCTGGGGAAGTCGCTGATTTTCGTGGAATTGATGCTTTTTGCGCCATAATCA
TCTTCATAGACCAACTGGCTGAACGTTTCCAAACCATGCAGGGCCCCCCAAACATTTGGTGCTTTCAGGACAGCAAATGGCTGATCCA
CTGACAGTTCATATGACTCATCCGCCGTGATACTGGGGTAGCCGTCACACTCCGAGTCAGGTGATGTGATCCACACCTGCAGCTGTGA
CAGCTCAGAAGGCCCCGTTTGCCTGTTTTTGTTCTGCCACCGTCTTCTATTGCCACCAAACATGTATTCATAATACCTCCTGTAGGCG
TTCTGCAGGACGTTGCAGCTCGGCCCAGCAGAGGACTCTCTCGCGTCAACGATGCTGAAGCTAGAACCGATTATCTTGAATGAAACCT
CGGAGCTCTGCACTTTCTGCGGGAGAGGCCACAGTGAGCTGGACTGTTGTCCTGAACTAAGGCCGAGGGTCAGGAGCAGCAGTGGCAC
AGCAAACATCAGCTCGGCCATCGCTCCGATTTACAGCCGGAGAGTCACCCCCCGTACTCTACACACGGCTCCCCCGTACTCT

>Sse_Gonadas_ClGo_7:

AGAGTACGGGGGGATTTCGGCTCAAAAACTTGTGGTTCAAGTCAGAGTACGGGGGATGACCCCTACCCATTTCGCCTTCTCATCAGCT
TTTATTCTTGGACTAACTGGGCTAGCATTCCACCGATCCCATCTGCTCTCCGCCCTCATTTGCCTAGAAGCAATAATAGTCTCGTTAT
TCATTGCCCTCTCTCTTTGAACAATACAGCTAAACTCTACTAACTTCTCCGCCTCCCCCCTCCTTCTACTAGCATTCTCAGCCTGCGA
AGCAAGCATAGGCTTAGCCCTCCTGGTCGCCACAGCCCGAACCCACGGCTCCGACCAGATCCGAAACTTAAACCTGTTACAATGCTAA
AAGTCCTAATTCCCACCCTAATGCTAGCCCCAACCATTTGGCTTTCAAAACCCAAATGACTCTGACCAACTACCTTAGCCCATAGCTT
TTTAATTGCACTTATTAGCCTAACTTGACTCCAAAACCTCTCGGAAAGTAGCTGAGCCTCATTAAGCCCTATTATAGCAACCGATTCT
ATCTCGACCCCCTTGCTTGTTCTATCCTGCTGACTCCTTCCTCTCATAATTTTAGCCAGTCAAAACCATATAGCCAAAGACCCCTCTA
ACTTTCAACGAGCCTACATCAGCCTCCTCACTTCCCTTCAGTTCTTTCTTATTTTAGCATTTGGTGCAACCGAACTAATTATGTTCTA
TGTAATATTTGAGGCTACCCTTATCCCGACCTTAATTATTATTACTCGCTGAGGAAACCAGGCAGAACGCCTAAACGCGGGAACCTAC
TTTTTATTTTATACCCTTGCAGGCTCCCTCCCACTACTCATTGCCCTACTATCCTTACAAGGCTGAACAGGAACAATGTCCCTCCTAA
CCCTCCCTTTTTCAGACCTTGCTCCATTAACCCCCTACACAAGCATACTTTGATGAGCAGCCTGCCTTCTTGCCTTCCTAGTAAAAAT
ACCATTATACGGCGTCCACCTGTGACTCCCAAAAGCCCATGTAGAAGCCCCCGTGGCCGGGTCCATGGTACTTGCCGCAGTTCTTCTA
AAACTCGGAGGATACGGAATGATACGCATCATAATTATACTGGAACCCCTCACCAAACAACTTTGCTACCCCTTCATTATTCTCGCCA

TATGGGGGATCATCATAACTGGCTCTATTTGTCTCCGACAAACAGATCTCAAATCTCTTATTGCCTATTCTTCAGTAAGCCATATGGG
TCTTGTAGCCGGGGGCATTTTAATTCAAACCCCCTGAGGATTTACAGGAGCCCTCATCCTCATAATTGCCCACGGATTAACCTCATCC
GCACTCTTTTGCCTTGCAAACACCAACTATGAACGAACCCACAGCCGAACTATAGTACTAGCCCGAGGCCTACAAGTGATTCTTCCCC
TGATAGCAGCTTGATGGTTCATTGCCAGCCTGGCAAATTTAGCCCTTCCTCCTCTCCCCAACCTTATAGGGGAATTAACAATCATTAC
CTCACTATTCAACTGGTCTTGATGAACCCTTATTCTGACAGGTGCAGGGACCCTAATTACAGCCGGCTATTCCCTGTCCATGTTCCTC
ACCACCCAACGAGGACCCCTCCCCCCCACACTATCCTGGCATTAGAACCCTCCCACACCCGAGAACACCTTCTCATCGCACTTCACCT
ACTCCCCCTCCTTCTCCTGA

>Sse_Gonadas_ClGo_8:

AGAGTACGGGGGAACAACAGTCGCAGCGTTCTACGTTTTTCTGTAGAAGGCTTGTAAACATTCTCCAGTCGGCGTCGCTGCTCATAGA
ACGTTCAACCTGTTCGCTGGGCGCTTAGTAGATATAAACGTCTACGTAGAGTACGGGGGCTCTTCGACCACCGCCATAGTGGGGAAAG
GACTGTGGACAAAATGTCGTCCCACAAGACATTCAGGATCAAGCGTTTCCTCGCTAAGAAGCAGAAACAGAACAGGCCGATTCCTCAG
TGGATCAGAATGAAGACTGGCAACAAGATCAGGTACAACTCCAAGAGGAGACACTGGAGGAGGACCAAGCTTGGCCTGTAAACATGAG
GGTCTCTGGATCCCTGTGTCCTCCCCAGACCTCCAGCTGCTGTGCCTGACTCTGGAGGGGGAGCGCAATGTTATCTTGTCATGTTTGA
TAATAAAAGATCTGTGAAT

AII.IV. Secuencias consenso de los clusters de nivel 1 seleccionados en el paquete sistema inmune

>Sse_Inmuno_ClIn_1:

GTCTCGACAGAATCGAAGTTGGCTGACCAGTGTACCGGTCTCCAGGGATTCCTCATCTTCCACTCCTTCGGTGGTGGTACCGGATCCG
GATTCACTCCCTCCTTATGGAACGTCTTCGTGATAGGAAAGAAATCCAAGCTGGAATTCTCCATCTACCCAGCTCCCCAGGTGCCACC
GCCGTGTTGAGCCATACAACTCCATCCTGACCACCCACACCACCCTGGAACACTCCGACTGCGCCTTCATGGTCGACAATGAGGCTAT
CTATGATATCTGCAGACGTAACTTGGACATTGAAAGACCAACCTACACCAACTTGAACAGGTTGATTGGTCAGATCGTCTCCTCCATC
ACGCCTCTCTCAGATTTGATGGTGCCCTCAACGTNGACCTGACAGAGTTCCAGACCAACTTGGTGCCATACCCTCGTATCCACTTCCC
TCTNGCCACCTATGCCCCAGTCATCTCNGCTGAGAAGGCCTACCATGAGCAGCTCTCAGTTGCTGAGATCACCAACGCCTGCTTTGAG
CCAGCCAATCAGATGGTGAAATGTGACCCTCGTCACGGCAAGTACATGGCCTGCTGCCTGTTGTACCGTGGTGATGTTGTCCCCAAAG
ATGTCAACGCCGCCATCGCAACCATCAAGACCAAGCGNACCATCCAGTTTGTGGACTGGTGTCCCACTGGTTTCAAGGTCGGCATCAA
CTACCAGCCACCAACTGTGGTTCCTGGTGGNGACCTGGCCAAGGTNCAGAGNGCNGTGTGCATGCTGAGCAACACCACTGCATCGCTG
AGGCCTGGGCTCGTCTCGACCACAAGTTTGACCTGATGTACGCCAAGAGAGCCTTTGTCCACTGGTATGTTGGAGAAGGTATGGAGGA
GGGAGAGTTCTCTGAGGCCAGAGAAGACATGGCCGCCTGGAGAAGGATTATGAAGAGGTCGGAGTTGACTCATTGAAGGGAGGGAGAG
GAGGAAGGAGAAGAATATTAAAGGACTCAACAACTCCTCAAACAAAATTCTGTTTTTGTTTATTTAATCTCCCATTTAGTGCTGAAAT

>Sse_Inmuno_ClIn_2:

GTATATCCACTCACCACCAACAAAAGAAGAATAGGCTATGAAAAGACGTCCCCGATGTGACTTCTACAGTGGACCGACATGGACGCTG
ATAATCTCGGACATATAGTTAATTTCGGCGCAGAGTATAAATAAAACTTTATACTAATACTACTACTAATTCTAATACTACTACTAAT
CATGGACCAATCTGGCCTACTGGCTTTGTGAGATTTAACCAATAAAACCATTAAAAACTTACCAGTTTAAAAGAGAATTGAACAAGTA
ATGTTCACCAAGTTTATGTACGGTGCCTCGCTTTGTACCTTATCGTTGGTCATGTGTGACTGGATCATATCATGGCTGAGTTCAGTGA
AAGTCATCTGACCTTGTTGATGTTATTGTAAGAAATTCAAATTTAATGTTTTAGGTATTTGAACTTCAGTGGAGTTCCTTTAAGAAAT
GTCTGTTTGTTTTGTTTGTTAAATCATGCAATTCCATGATATACTCCACTATATGTTTTTTCG

>Sse_Inmuno_ClIn_3:

TATCAAAAAATTTGACTTCCGCTTCTTTTCTACGCGGAGTTTTCCCAAGTGGATACATCATTAAGGACGTCCGTTTCTCAACATGGCT
GGTTCAGATACGGAAGGAGATGAAATTTCCACTGGGGCGGGAGCCATGGATACCATGACCGCCCTTCAAGAAGTCCTCAAAACTGCAC
TGATCCATGATGGTCTAGCACGTGGATTACATGAGTGTGCCAAGGCACTTGATAAACGTCAAGCTCATTTATGTGTACTTGCGAGCAA
TTGTGATGAAGCTATGTATTCAAAGTTGGTTGAAGCTCTTTGTGCTGAACATGGTATTAACCTCATTAAGGTT

>Sse_Inmuno_ClIn_4:

TATGACACATTATAGACACTAAAATTTACAGAGATCCGGATTAAATGAAAATCTTTAAAATGAAGGTCATTTAGGCTCTCCTCTTCCT
CCTCATCAAACTCGCCCTCTTCTCGCGTGGCTCTGGTACTGCTGGTATCGAACCAGTCGTTCATGTTTCTGCCTCGTAACTCCATCTC
GTCCATCCTCCCGTTACCATGAGAAGCCTTCGCGGAACATGCGGTGAACTGTCGAGATACGCTTGAACAGCTCCTGGATGGCGTGCTG
TTGCCAATGAAGGTGGCGGNCATCTTTAGNCCACGTGGTGGGATNTCACANACGGCTGTCTTGACGTTGTTGGGGATCCATTCAACAA
AGTAGCTGCTGTTCTTGTTCTGGACATTCANCATCTGTTCATCGACCTCCTTCATGGACATGCGGCCACGGAANATGGCAGCNACTGT
CAGGTANCGNCCGTGACGTGGGTCACAGGCGGCCATCATGTTCTTGGNGTCNAACATNTGCTGAGTNAGCTCAGGGACNGTGAGTGCC
TGTATTGAGTTGATGGACTGAAATGTGGCGTTATATGGTTCTACAACTGTGTCTGAGACCTTGGGAGGACAACGAGAAGGCATGATTC
GTCGGGTATCTTCACGGATTTGCTGATGAGAGGTTCCCATACCGAGCCGGTACCACCACCAAGAGAGTGTGTGAGTTGGAACCCTGGA
GGCAATCACAGCTTTCTGATTCTTTCACAACATCAGGACGAGTCACCAGCTCGCCCCTCGTGTAGTGACCTTTGGCCCAGTTGTTTCC
AGCCCACTCTGCCAAACAAAGTTGTCTGGTCTAAATGACCAAAGGTCCGACTCACAGATCCATGGTTCCAGGCTCCAGTCGACCAGAC
TGCACGGGGACATATTTCCCCAGAGCTTCATTATAGTAGACTTGATTCTCTCAAGTTGAAGGTCGATCCCCGTGGTATGTACCGGTGG
GGTCAATTCCATGCTCTCTGAGATACTTCCCAGAACTTGCCCATTTGATTTCCGCATGTCCGCTGAAGATGTACATTTCTCTCATTTG
TGTTTATAAACCTCTTTTTTAAATCCTTGTGAAG

>Sse_Inmuno_ClIn_5:

TTTCAAGGTTCGCAGTGTGTCGGAAGGATTGATTTACAGCTAAAGAAATGGCGTCTTCTAAAGTAACGTTAGTACAGGCCTTGATTTG
TGCGCTTTGTTGCCTACTTCTGTCACATCTGGCAAGTTCACAGGATCTACCAATGATACTAGTACTACAATAAAGTCAACAACACAAA
CGGCTTCAAACAAAAACAAGTGACCACTATGCGACAAATCGACTACATCCGGTGCTGGA

>Sse_Inmuno_ClIn_6:

ATTAAAAGATAGAGCAGCATTACTTTACTTGGACATTTTTTGGTAATCGAATTAAACAAGGATGAGGGCTTATATCTTATTTCCTTTG
TGTTATTTCCGCCATGTTGATGACGTCAGAAGCAATCGAGTGTCGTTAAGTGCAGAGCTGAGTGTATGGATGCCTACAACGTGTGCGC
AGGTGCCTGCAAAAATGGTT

>Sse_Inmuno_ClIn_7:

GACTGCAGTCGACTCGACCCTTGTCTGTATCATCTGTAGCCTGCAAAGCAACCACACCAGAGCTCAAAATGCCACTCACTGATATACT
GGAACAGGGTAAAATCACTGCAGCCCTGGAAGAGTGCAAAAATGCTGACTCTTTCCACCACAAGCAGTTCTTCAAGACTTGCGGTCTG
ACCGGAAAGAGTCATGCTGATGTGAAGAAGGTCTTCGACATCATTGACCAGGACCGGAGTGGTTTCGTTGAGGAGGACGAGCTGAAGC
TGTTCCTGCAGAACTTCAAGAAAGATGCACGTGCA

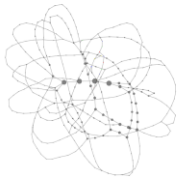
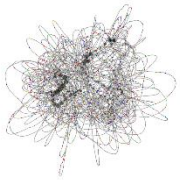
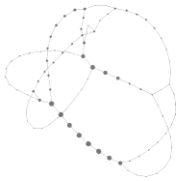
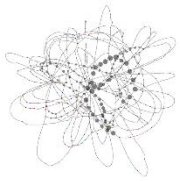
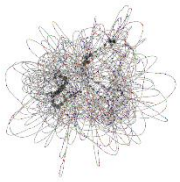
>Sse_Inmuno_ClIn_8:

TCGCGTTCAAACAAGAAGGAAACGTTGCACTCAACACATCTCGCACAAAGCCGCGTGACTTCCAAATCCCGTTCCTCTTTCAACTCTC
CTCCAAGATGCCCACCAAGAAGACCAAGACCAGGAAGCTACGTGGACACGTCAGCCACGGGCATGGTCGTATTGGCAAGCACAGAAAG
CATCCTGGAGGTCGTGGTAATGCAGGTGGTATGCATCACCACAGAATCAACTTCGACAAATACCATCCAGGCTACTTTGGTAAAGTGG
GTATGAGACATTACCACCTGAAAAGAACACAACTCAT

AIII. RESULTADOS DEL ANÁLISIS DE LOS *K-MEROS*

AIII.I. *K-meros* del paquete S0

Tabla III.I.I. Información de los parámetros de los k-meros más influyentes en la construcción del cluster ClS0_1.

Variante	*K-meros*	*Score*	Longitud del monómero	Grafo
ClS0_1_1	15	0.4871	38	
CLS0_1_2	12	0.4742	76	
CLS0_1_3	1	0.4586	38	
CLS0_1_4	3	0.4124	38	
CLS0_1_5	5	0.3893	38	

Secuencias de los *k-meros*:

>ClS0_1_1:

AAATGTGACAAAAAAGTCATAGTTTAGTATGTCGTCCA

>ClS0_1_2:

AAAATGTGACAAAAAAGTCATAGTTTAGTATGTCGTCCAAAATTTGACAAAAAAGTCATAGTATAGTATGTCGTCC

>ClS0_1_3:

TGTGACAAAAAAGTCATAGTTTAGTATGTCGTCCAAAA

>ClS0_1_4:

GTATGTCGTCCAAAATCTGACAAAAAAGTCATAGTTTA

>ClS0_1_5:

AAAATCTGACAAAAAAGTCATAGTATAGTATGTCGTCC

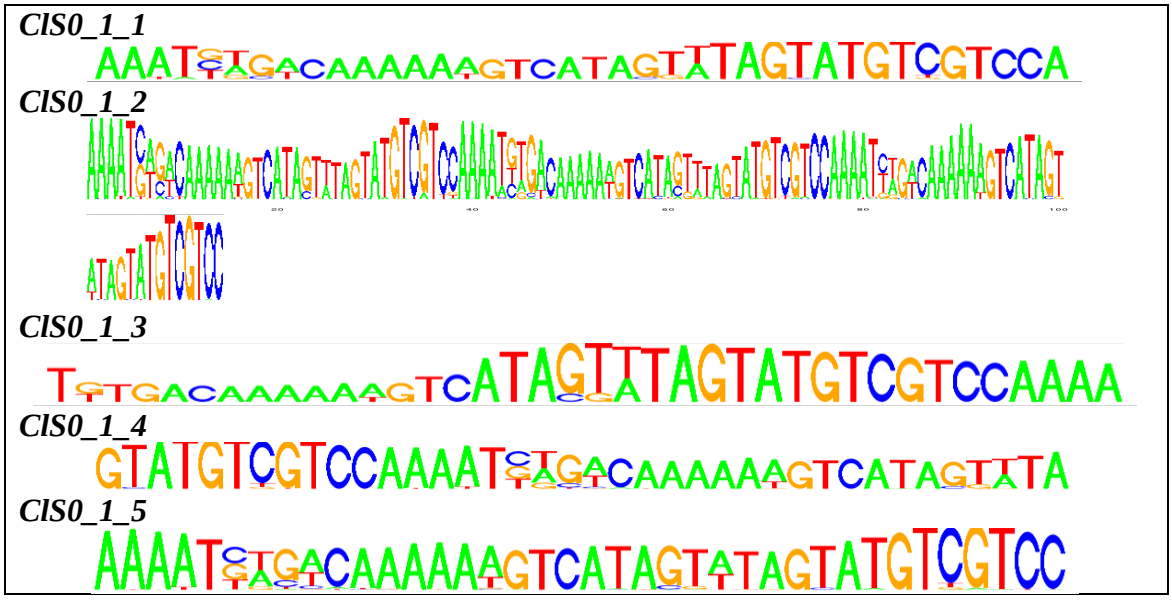
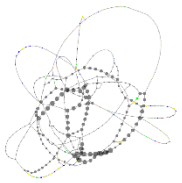
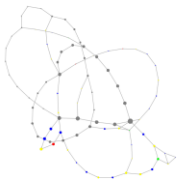
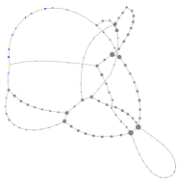
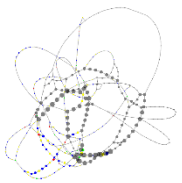


Figura III.I.I Logos de los k-meros más influyentes en la construcción del cluster ClS0_1.

Tabla III.I.II. Información de los parámetros de los k-meros más influyentes en la construcción del cluster ClS0_2.

Variante	*K-meros*	*Score*	Longitud del monómero	Grafo
ClS0_2_1	11	0.6304	21	
CLS0_2_2	23	0.4614	42	
CLS0_2_3	11	0.4026	42	
CLS0_2_4	19	0.3796	21	
CLS0_2_5	23	0.3710	84	

Secuencias de los *k-meros*:

>ClS0_2_1:

TTGTAGTTGTTGTGGCTGGTG

>ClS0_2_2:

GGTGTCGTAGTTGTTGTTGCTGGTGTTGTAGTTGTTGTGGCT

>ClS0_2_3:

GTAGTTGTTGTGGCTGGTGTCGTAGTTGTGGTGGCTGCTGTT

>ClS0_2_4:

GTTGTAGTTGTTGTGGCTGGT

>ClS0_2_5:

TCGTAGTTGTGGTGGCTGGTGTTGTAGTTGTTGTTGCTGGTGTTGTAGTTGTTGTGGCTGGTGTCGTAGTTGTGGTGGCTGGTG

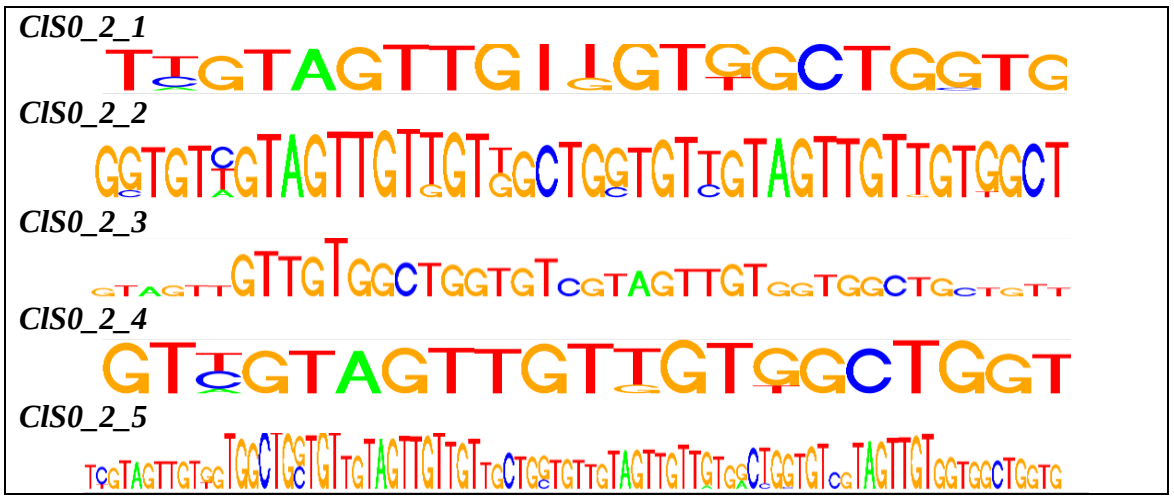
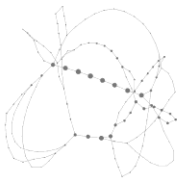
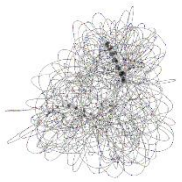
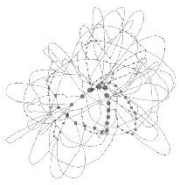
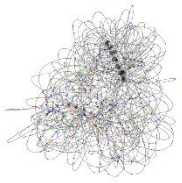


Figura III.I.II Logos de los k-meros más influyentes en la construcción del cluster ClS0_2.

AIII.II. *K-meros* del paquete S4

Tabla III.II.I. Información de los parámetros de los k-meros más influyentes en la construcción del cluster ClS4_1

Variante	*K-meros*	*Score*	Longitud del monómero	Grafo
ClS4_1_1	11	0.5639	38	
CLS4_1_2	15	0.4953	38	
CLS4_1_3	23	0.4894	76	
CLS4_1_4	19	0.4285	38	
CLS4_1_5	23	0.4059	38	

Secuencias de los *k-meros*:

>ClS4_1_1:

ACATACTAAACTATGACTTTTTTGTCAGATTTTGGACG

>ClS4_1_2:

ATACTATGACTTTTTTGTCAGATTTTGGACGACATACT

>ClS4_1_3:

CTTTTTTGTCAGATTTTGGACGACATACTAAACTATGACTTTTTTGTCACATTTTGGACGACATACTATACTATGA

>ClS4_1_4:

TGACTTTTTTGTCTGATTTTGGACGACATACTATACTA

>ClS4_1_5:

CTTTTTTGACTGATTTTGGACGACATACTATACTATGA

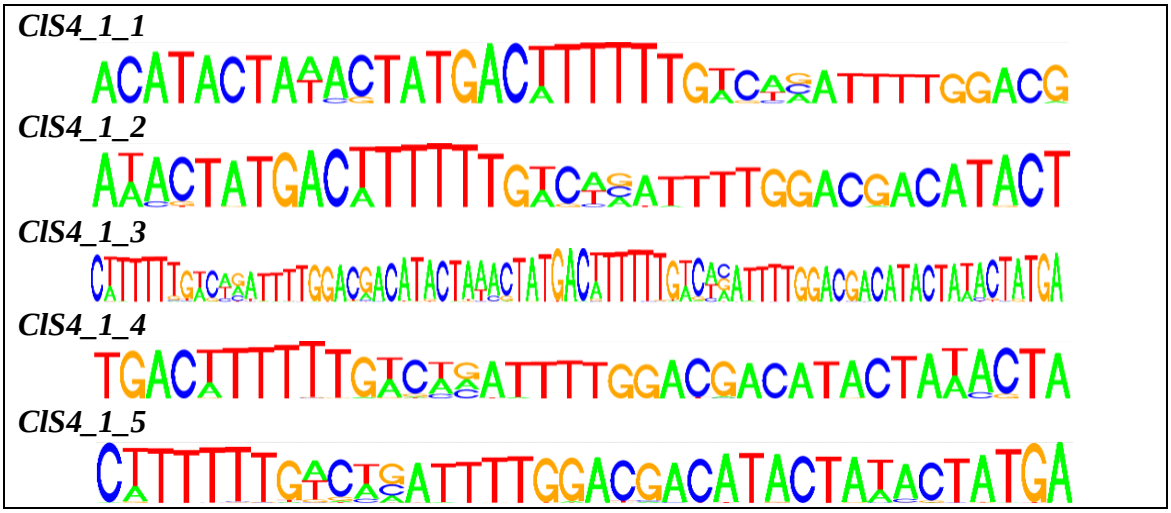
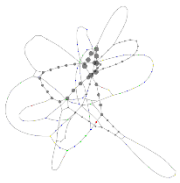
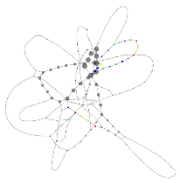
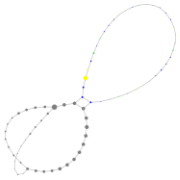
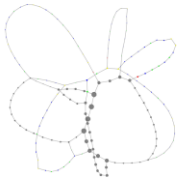
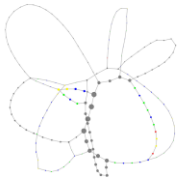


Figura III.II.I. Logos de los k-meros más influyentes en la construcción del cluster ClS4_1.

Tabla III.II.II. Información de los parámetros de los k-meros más influyentes en la construcción del cluster ClS4_2.

Variante	*K-meros*	*Score*	Longitud del monómero	Grafo
ClS4_2_1	15	0.412	27	
CLS4_2_2	15	0.385	45	
CLS4_2_3	11	0.379	27	
CLS4_2_4	19	0.333	27	
CLS4_2_5	19	0.323	45	

Secuencias de los *k-meros*:

>ClS4_2_1:

GCTCCTCAGGATTCTCTGGTTCTTCTG

>ClS4_2_2:

ATTCTCTGGTTCTTCTGGCTCCTCTGGTTCTTTTGGCTCCTCAGG

>ClS4_2_3:

ATTCTCTGGTTCTTTTGGCTCCTCAGG

>ClS4_2_4:

TCTTCTGGCTCCTCAGGATTCTCTGGT

>ClS4_2_5:

CTGGATCCTCTGGTTCTTTTGGCTCCTCAGGATTCTCTGGTTCTT

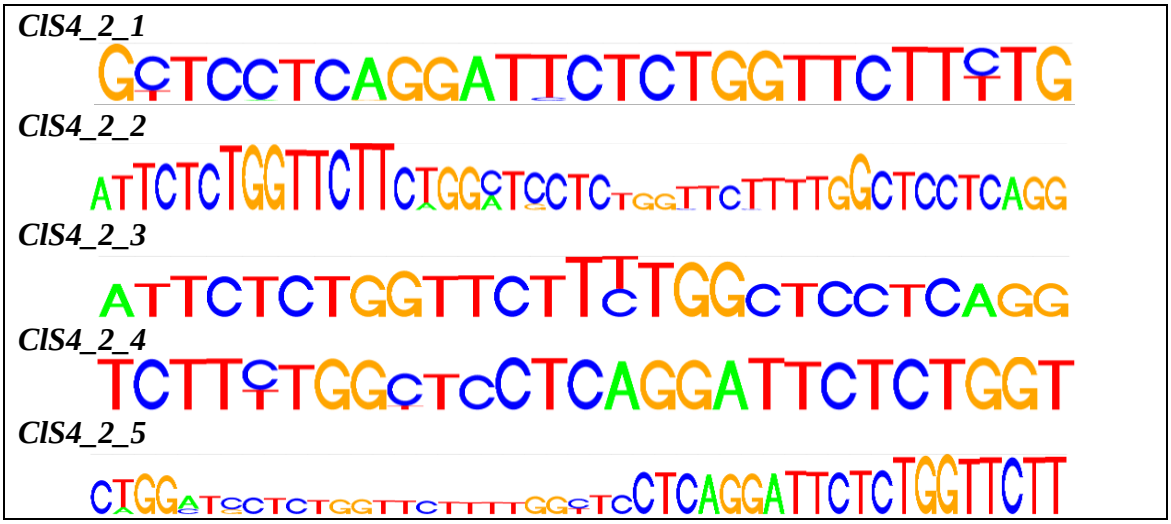
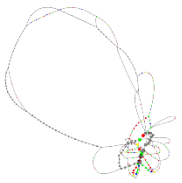


Figura III.II.II. Logos de los k-meros más influyentes en la construcción del cluster ClS4_2.

Tabla III.II.III. Información de los parámetros de los k-meros más influyentes en la construcción del cluster ClS4_3.

Variante	*K-meros*	*Score*	Longitud del monómero	Grafo
ClS4_3_1	11	0.454	138	
CLS4_3_2	11	0.433	165	
CLS4_3_3	11	0.419	150	
CLS4_3_4	11	0.414	27	
CLS4_3_5	15	0.342	27	

Secuencias de los *k-meros*:

>ClS4_3_1:

CACAGGTCCTGTCGTTACAACTGTGGTTATGACCTGGGAGTCTGCTCCTGCCGCAGCTCGTGTAGATACAATGGTACCTGTTGCTATG
ACTACAACAACTACTGCCTTGGAACGACAGCACCACCAGAAACAACAGGA

>ClS4_3_2:

CAACCACAGGACACAGGTCCTGTCGTTACAACTGTGGTTATGACCTGGGAGTCTGCTCCTGCCGCAGCTCATGTAACTACTATGGTAC
CTGTTGCTATGACTACAACAACTACTGCCTTGGAACGACAGCACCTCCAGAAACAACAGCAACGACTGCACCACCAA

>ClS4_3_3:

CAACAACAGCAACGACTGCACCACCAGAAACAACAGGACACAGGTCCTGTCGTTACAACTGTGGTTATGACCTGGGAGTCTGCTCCTG
CCGCAGCTCATGTAACTACTATGGTACCTGTTGCTATGACTACAACAACTACTGCCCTGGAA

>ClS4_3_4:

CACCTGAAACAACAGCAACGACTGCAC

>ClS4_3_5:

CTGAAACAACAGCAACGACTGCACCCC

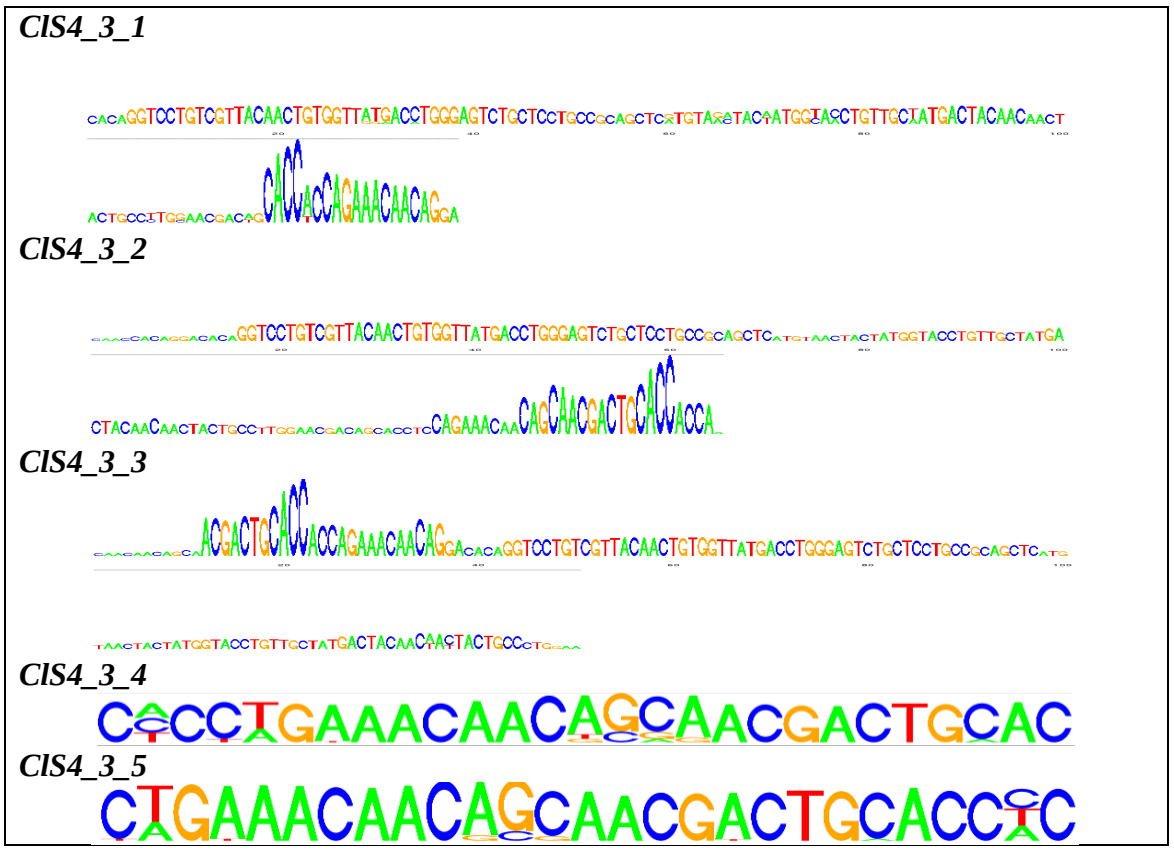
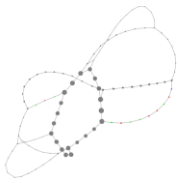
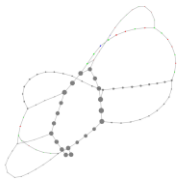
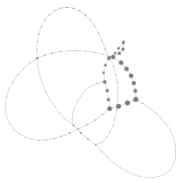
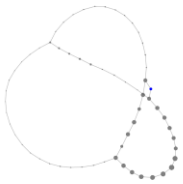
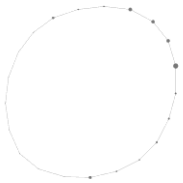


Figura III.II.III. Logos de los k-meros más influyentes en la construcción del cluster ClS4_3.

Tabla III.II.IV. Información de los parámetros de los k-meros más influyentes en la construcción del cluster ClS4_4.

Variante	*K-meros*	*Score*	Longitud del monómero	Grafo
ClS4_4_1	15	0.72	42	
CLS4_4_2	15	0.69	21	
CLS4_4_3	23	0.64	42	
CLS4_4_4	19	0.54	42	
CLS4_4_5	11	0.54	21	

Secuencias de los *k-meros*:

>ClS4_4_1:

TACTACAACCACAACTCCAACTACTACAACCACAACTCCCAC

>ClS4_4_2:

CCACTACTACAACCACAACTC

>ClS4_4_3:

CAACCACAACTCCCACCACTACAACCACAACTCCCACTACTA

>ClS4_4_4:

CCCAACTACTACAACCACAACTCCCACTACTACAACCACAAC

>ClS4_4_5:

ACTACAACCACAACTCCCACT

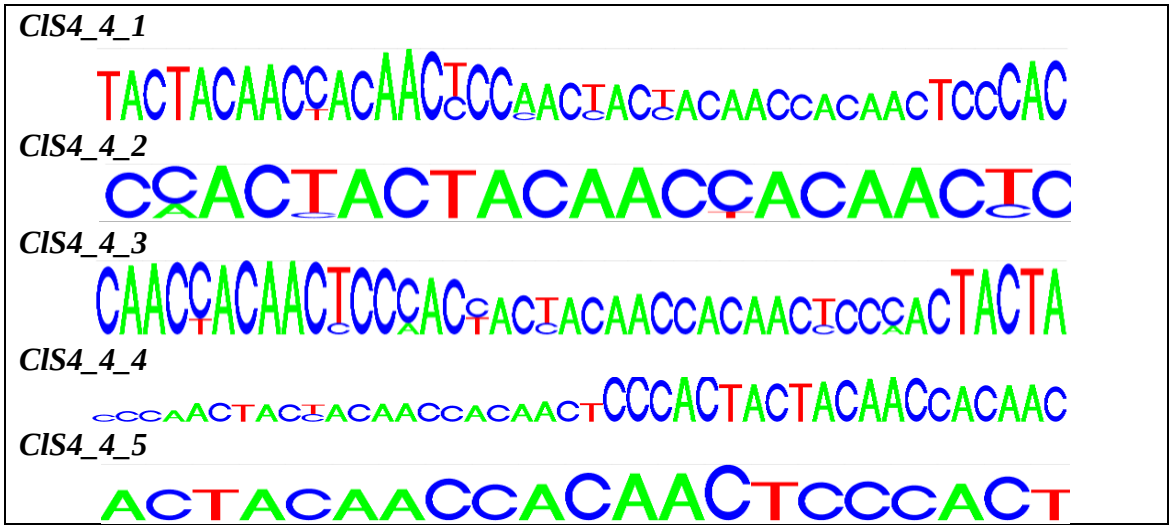
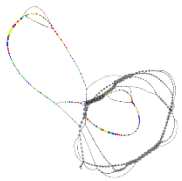
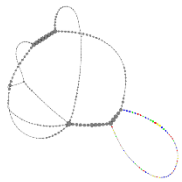


Figura III.II.IV. Logos de los k-meros más influyentes en la construcción del cluster ClS4_4.

Tabla III.II.V. Información de los parámetros de los k-meros más influyentes en la construcción del cluster ClS4_5.

Variante	*K-meros*	*Score*	Longitud del monómero	Grafo
ClS4_5_1	15	0.578	102	
CLS4_5_2	11	0.504	89	
CLS4_5_3	11	0.471	102	
CLS4_5_4	19	0.437	102	
CLS4_5_5	27	0.430	102	

Secuencias de los *k-meros*:

>ClS4_5_1:

TACTACAACCACAACTCCAACTACTACAACCACAACTCCCAC

>ClS4_5_2:

CCACTACTACAACCACAACTC

>ClS4_5_3:

CAACCACAACTCCCACCACTACAACCACAACTCCCACTACTA

>ClS4_5_4:

CCCAACTACTACAACCACAACTCCCACTACTACAACCACAAC

>ClS4_5_5:

ACTACAACCACAACTCCCACT

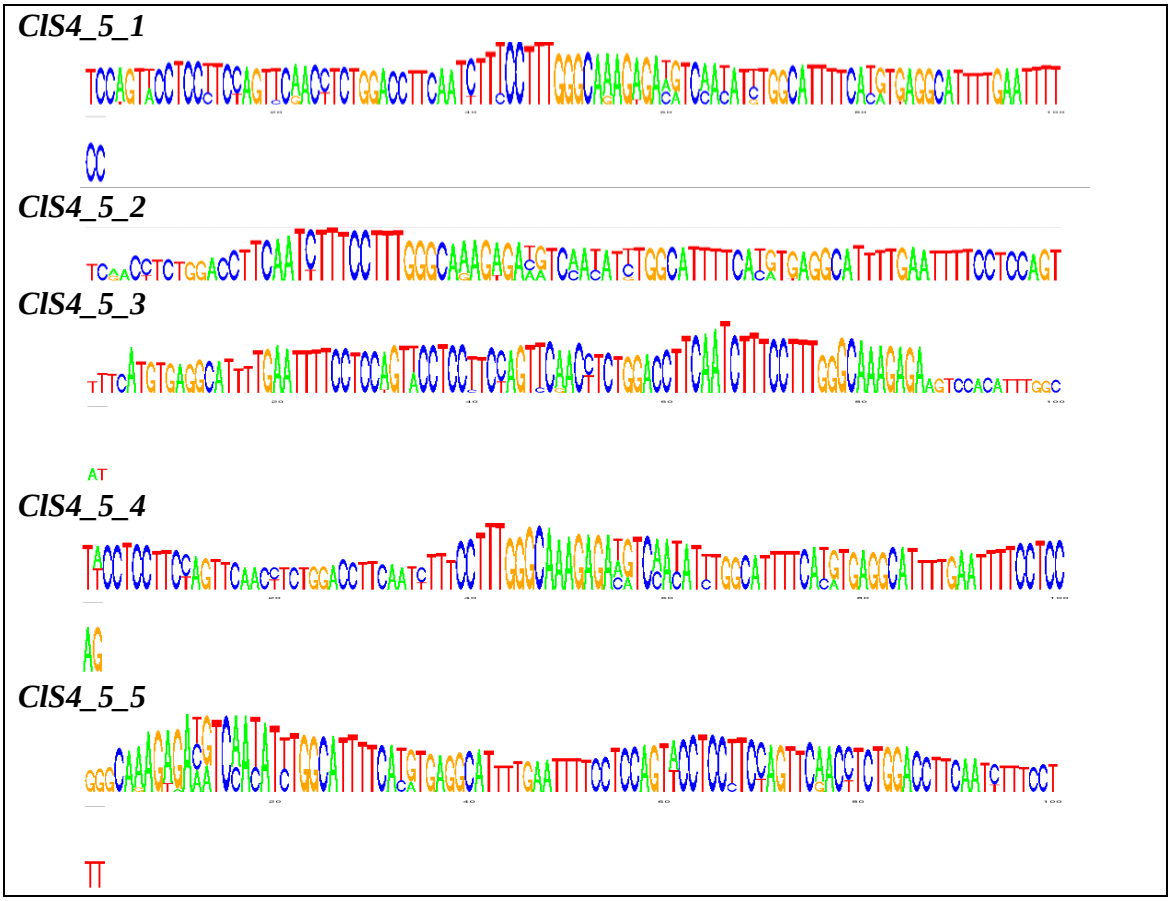
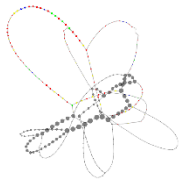
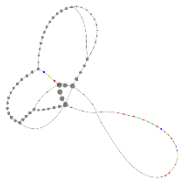
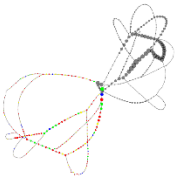
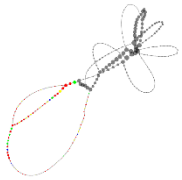
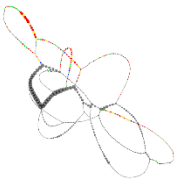


Figura III.II.V. Logos de los k-meros más influyentes en la construcción del cluster ClS4_5.

Tabla III.II.VI. Información de los parámetros de los k-meros más influyentes en la construcción del cluster ClS4_6.

Variante	*K-meros*	*Score*	Longitud del monómero	Grafo
ClS4_6_1	15	0.45	57	
CLS4_6_2	11	0.41	57	
CLS4_6_3	23	0.37	57	
CLS4_6_4	19	0.37	57	
CLS4_6_5	27	0.31	144	

Secuencias de los *k-meros*:

>ClS4_6_1:

GTCAAGAAACCATCTAAAGAAGAGAAAGAGGTGAAGCCTACCAAAGAAAAGAAAGAA

>ClS4_6_2:

TCAAGAAACCATCTAAAGAAGAGAAAGAGGTGAAGCCTACCAAAGAAAAGAAAGAAG

>ClS4_6_3:

AAGAGGTGAAGCCTACCAAAGAAAAGAAAGAAGTCAAGAAACCATCTAAAGAAGAGA

>ClS4_6_4:

GTCAAGAAACCATCTAAAGAAGAGAAAGAGGTGAAGCCTACCAAAGAAAAGAAAGAA

>ClS4_6_5:

ACCATCTAAAGAGAAAGAGGCAGAGAAACCTCCGAAAGAAGAAAAAGAGGTTAAA

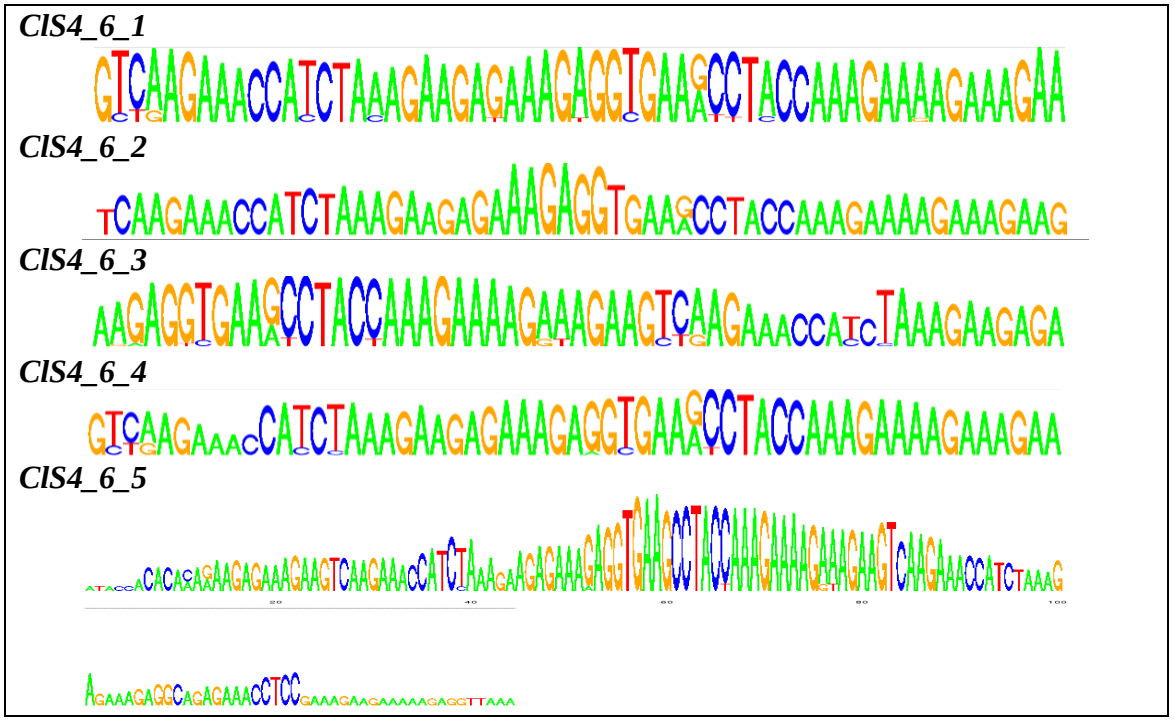
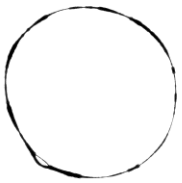
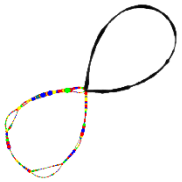
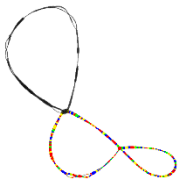


Figura III.II.VI. Logos de los k-meros más influyentes en la construcción del cluster ClS4_6.

Tabla III.II.VII. Información de los parámetros de los k-meros más influyentes en la construcción del cluster ClS4_7.

Variante	*K-meros*	*Score*	Longitud del monómero	Grafo
ClS4_7_1	19	0.561	702	
CLS4_7_2	23	0.559	702	
CLS4_7_3	27	0.510	702	
CLS4_7_4	19	0.331	378	
CLS4_7_5	15	0.301	324	

Secuencias de los *k-meros*:

>ClS4_7_1:

CCGAGTGTTGACGTCAAAGCTCCAGAGATTGACATCGAAGCTCCTGATGTAAAACTCCATGGACCAAATATCAAATTACCATCAATTT
CAGCGCCCAAGCTCCCAGACTGGGATCTTAAACTGAAAGGGCCCAAAGTAAAGGGAGATGTTGATGTCTCAGTTCCAAAGATTGAGGG
TGATATAAAAGGACCCAAACTTGATATTGAAGGACCAGATGTTGACCTTGATGGTAAAACAGGAGGATTTAAAATGCCTAAATTCAAA
ATGCCATCCTTTGGATTTAAAGGCTCACATGGTGAAGGGCCAGAGGTTGATGTTAGTCTCCCGGAGGCTGATATTGATATCAGAGCTC
CAGATATTGATATCAAAGGACCAGAGGTTGACCTGGAAAGTCCCAGTGGAAAGATCAAGGGATCAAAATTCAAAATGCCAAACATCAA
AGGACCTCAAATCTCTATGCCTGATGTGGATTTTAATTTGAAAGGTCCAAACTGGAAAGGCGGTGTGGATGTTTCAGGTCCAAAGATT
AAAGGAGACATAGGAAAACCTGACATTGATTTCAAAGGTCCAGGGATTGATATTGAAGGACCAAAGGCTGGATTTGAAATGCCTAAAA
TCAAAATGCCAACTTTCAAAGGTGCTAAAATGGAGGGCCCAGATATTGATGTGAACCTCCCTAAGGCTGACTTTGATGTGAACTTA

>ClS4_7_2:

AGTGTTGACGTCAAAGCTCCAGAGATTGACATCGAAGCTCCTGATGTAAAACTCCATGGACCAAATATCAAATTACCATCAATTTCAG
CGCCCAAGCTCCCAGACTGGGATCTTAAACTGAAAGGGCCCAAAGTAAAGGGAGATGTTGATGTCTCAGTTCCAAAGATTGAGGGTGA
TATAAAAGGACCCAAACTTGATATTGAAGGACCAGATGTTGACCTTGATGGTAAAACAGGAGGATTTAAAATGCCTAAATTCAAAATG
CCATCCTTTGGATTTAAAGGCTCACATGGTGAAGGGCCAGAGGTTGATGTTAGTCTCCCGGAGGCTGATATTGATATCAGAGCTCCAG
ATATTGATATCAAAGGACCAGAGGTTGACCTGGAAAGTCCCAGTGGAAAGATCAAGGGATCAAAATTCAAAATGCCAAACATCAAAGG
ACCTCAAATCTCTATGCCTGATGTGGATTTTAATTTGAAAGGTCCAAACTGGAAAGGCGGTGTGGATGTTTCAGGTCCAAAGATTAAA
GGAGACATAGGAAAACCTGACATTGATTTCAAAGGTCCAGGGATTGATATTGAAGGACCAAAGGCTGGATTTGAAATGCCTAAAATCA
AAATGCCAACTTTCAAAGGTGCTAAAATGGAGGGCCCAGATATTGATGTGAACCTCCCTAAGGCTGACTTTGATGTGAACTTACCG

>ClS4_7_3:

CCGAGTGTTGACGTCAAAGCTCCAGAGATTGACATCGAAGCTCCTGATGTAAAACTCCATGGACCAAATATCAAATTACCATCAATTT
CAGCGCCCAAGCTCCCAGACTGGGATCTTAAACTGAAAGGGCCCAAAGTAAAGGGAGATGTTGATGTCTCAGTTCCAAAGATTGAGGG
TGATATAAAAGGACCCAAACTTGATATTGAAGGACCAGATGTTGACCTTGATGGTAAAACAGGAGGATTTAAAATGCCTAAATTCAAA
ATGCCATCCTTTGGATTTAAAGGCTCACATGGTGAAGGGCCAGAGGTTGATGTTAGTCTCCCGGAGGCTGATATTGATATCAGAGCTC
CAGATATTGATATCAAAGGACCAGAGGTTGACCTGGAAAGTCCCAGTGGAAAGATCAAGGGATCAAAATTCAAAATGCCAAACATCAA
AGGACCTCAAATCTCTATGCCTGATGTGGATTTTAATTTGAAAGGTCCAAACTGGAAAGGCGGTGTGGATGTTTCAGGTCCAAAGATT
AAAGGAGACATAGGAAAACCTGACATTGATTTCAAAGGTCCAGGGATTGATATTGAAGGACCAAAGGCTGGATTTGAAATGCCTAAAA
TCAAAATGCCAACTTTCAAAGGTGCTAAAATGGAGGGCCCAGATATTGATGTGAACCTCCCTAAGGCTGACTTTGATGTGAACTTA

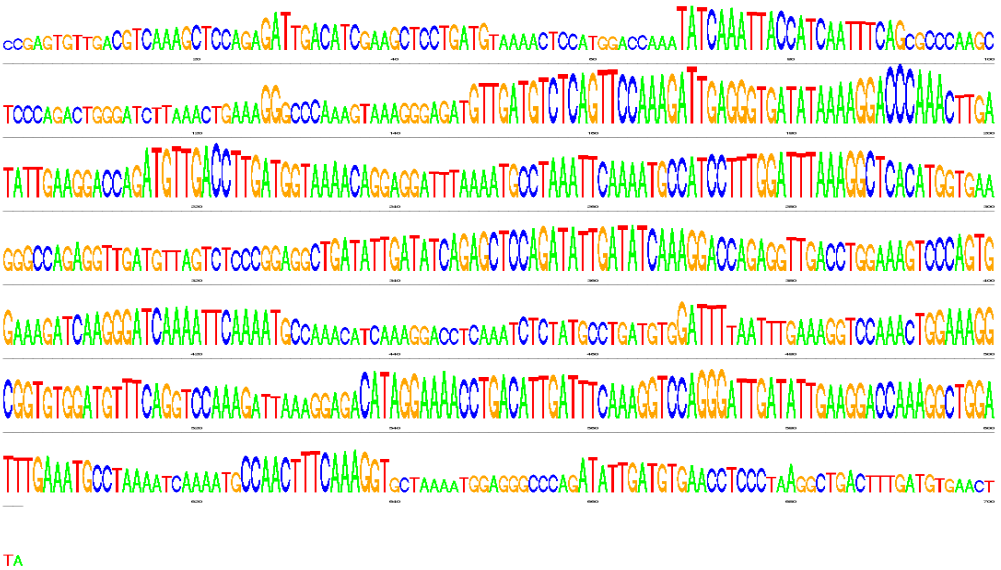
>ClS4_7_4:

TCAAATCTCTATGCCTGATGTGGATTTTAATTTGAAAGGTCCAAACTGGAAAGGCGGTGTGGATGTTTCAGGTCCAAAGATTAAAGGA
GACATAGGAAAACCTGACATTGATTTCAAAGGTCCAGGGATTGATATTGAAGGACCAGATGTTGACCTTGATGGTAAAACAGGAGGAT
TTAAAATGCCTAAATTCAAAATGCCATCCTTTGGATTTAAAGGCTCACATGGTGAAGGGCCAGAGGTTGATGTTAGTCTCCCGGAGGC
TGATATTGATATCAGAGCTCCAGATATTGATATCAAAGGACCAGAGGTTGACCTGGAAAGTCCCAGTGGAAAGATCAAGGGATCAAAA
TTCAAAATGCCAAACATCAAAGGACC

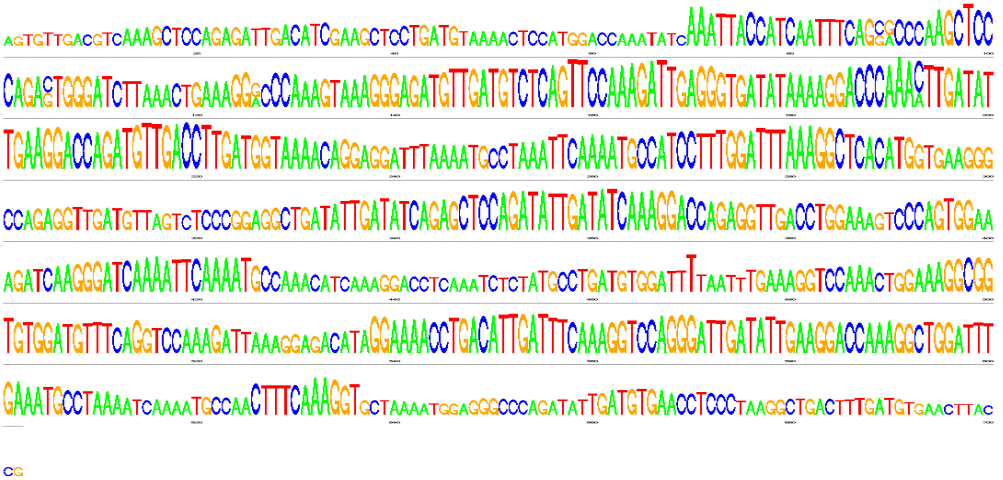
>ClS4_7_5:

TGGACCAAATATCAAATTACCATCAATTTCAGGACCCAAGCTCCCAGACTGGGATCTTAAACTGAAAGGGCCCAAAGTAAAGGGAGAT
GTTGATGTCTCAGTTCCAAAGATTGAGGGTGATATAAAAGGACCCAAACTTGATATTGAAGGACCAAAGGCTGGATTTGAAATGCCTA
AAATCAAAATGCCAACTTTCAAAGGTGCTAAAATGGAGGGCCCAGATATTGATGTGAACCTCCCTAAGGCTGACTTTGATGTGAACTT
ACCGAGTGTTGACGTCAAAGCTCCAGAGATTGACATCGAAGCTCCTGATGTAAAACTCCA

ClS4_7_1



ClS4_7_2



ClS4_7_3

ClS4_7_4

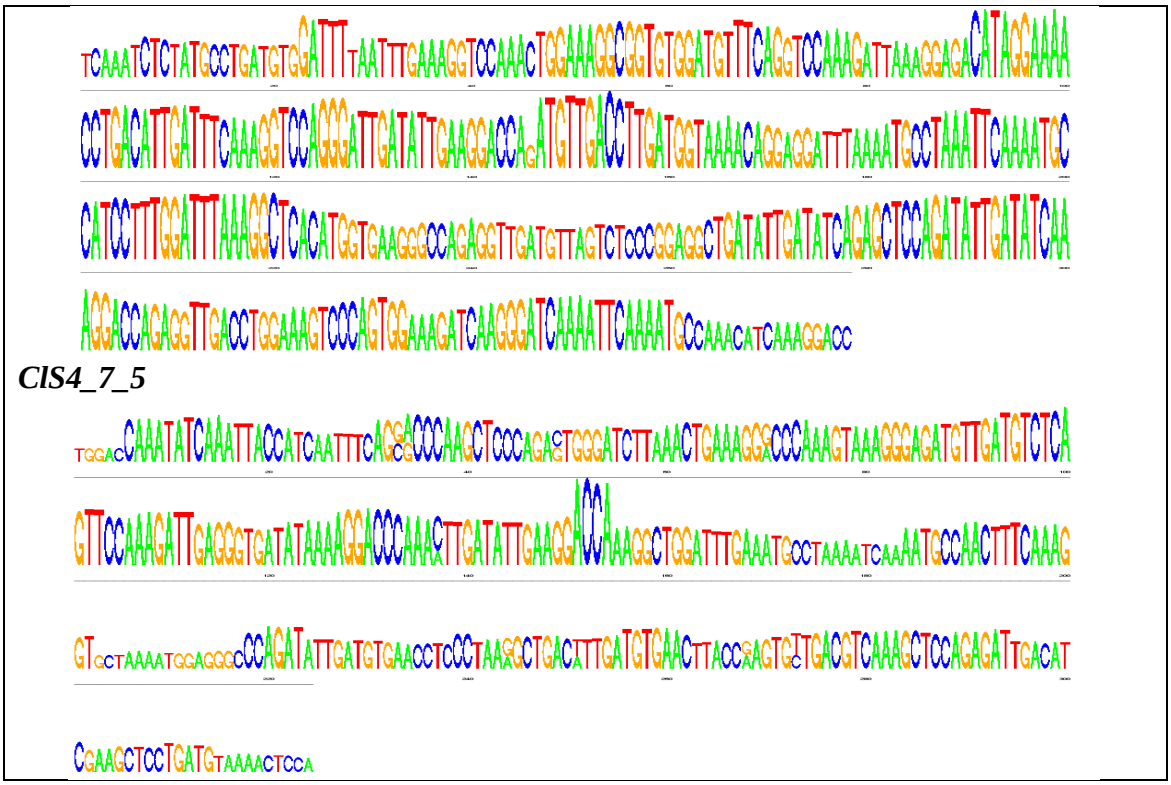
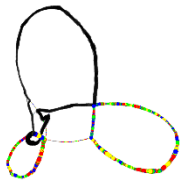


Figura III.II.VII. Logos de los k-meros más influyentes en la construcción del cluster ClS4_7.

Tabla III.II.VIII. Información de los parámetros de los k-meros más influyentes en la construcción del cluster ClS4_8.

Variante	*K-meros*	*Score*	Longitud del monómero	Grafo
ClS4_8_1	15	0.976	1164	
CLS4_8_2	19	0.965	1164	
CLS4_8_3	23	0.959	1164	
CLS4_8_4	27	0.937	1164	
CLS4_8_5	11	0.492	567	

Secuencias de los *k-meros*:

>ClS4_8_1:

CTTTGGCCTGAGGGTGCGTTTCGATGGTAACCACCATGCTGACGTCACCTTGCCGACCTCCTACAGTGGCCTGCTCTGTGGCATGTGT
GGAAATTTCAATAACAATCCAAGAGACGACAACCTGAAACCAGACCAAACACCAGCCGCTAACACCAATGAGTTGGGAGACAGCTGGC
AGGTTCCTGACCCGCGGCCTGACTGCACCAACGGTGGAGGACATGAAGAGTGTGACAAGAATGTGGAGGAGGAGGCCCAGAAACCAAC
CAGCTGTGGCATGATCACCGATCCTAACGGTATCTTCAAGCCTTGCCACTCTGTCGTGCCCCCGAACCAATACTTTGAAAACTGTGTG
TACGACGTGTGTGCAAATGGAGGTCAGACTGAGGCTCTGTGCCAGGCCATAGAGAGCTACGCTGATTTGTGTGCTGCAGCAGGAGTCC
CCATCGCATGGAGGAAAAACAACACCTTCTGTCCTATCAAGTGTCCCTCAGGCAGTCAGTACAATCCATGTACCTCTGCGTGTCCTCA
GCCCAGTTGCCAGGACCCTGCAGGCTCCGGTGGCTCCTGTAATCAGCCCTGTGTGGAGGGATGTGTCTGTAATCCTGGGCTCATCCTC
AGTGGAGACAAATGTGTCCCGCTCAGTGAGTGTGGATGCACTGATGAAGGTGGAAATTACAGGCCGACAGGAGACACTTGGTTCTCAG
AGAAGGACTGTTCAGAGCGCTGTAAGTGTAACGGCAACCACAACATCACCTGCGAGCCATGGCAGTGCAGCCCTACACAGGAGTGTAA
GGTGGTGGAGGGAGTACTAGGCTGTTACTCTAGAGGAAATGGAATCTGCTCAGTATCTGGTGATCCTCACTACAACACCTTTGACAAA
GTAACCCACCACTACATGGGGTCTTGCTCCTACACCTTGACCAAACCCTGCAACGTCTCCACCGACTTGCCGTACTTCACCGTGGACA
CCCAGAACGAGCACAGAGGAAGTAACAAGAGGGTTTCCTATGTCAGAGCTGTAGTGATCAATGTGAGCGGTGTGACTGTCATCCTTGG
CAAGGGACGCAAAGTCCAGGTCAATGGGACAGCAGTCGTCCCACCTTTGAATCCTGCCAAAGGAGTCAAGATCTACTTAAGTGGAAAG
TTTGTCGTCCTGGAGACAGA

>ClS4_8_2:

TTGGCCTGAGGGTGCGTTTCGATGGTAACCACCATGCTGACGTCACCTTGCCGACCTCCTACAGTGGCCTGCTCTGTGGCATGTGTGG
AAATTTCAATAACAATCCAAGAGACGACAACCTGAAACCAGACCAAACACCAGCCGCTAACACCAATGAGTTGGGAGACAGCTGGCAG
GTTCCTGACCCGCGGCCTGACTGCACCAACGGTGGAGGACATGAAGAGTGTGACAAGAATGTGGAGGAGGAGGCCCAGAAACCAACCA
GCTGTGGCATGATCACCGATCCTAACGGTATCTTCAAGCCTTGCCACTCTGTCGTGCCCCCGAACCAATACTTTGAAAACTGTGTGTA
CGACGTGTGTGCAAATGGAGGTCAGACTGAGGCTCTGTGCCAGGCCATAGAGAGCTACGCTGATTTGTGTGCTGCAGCAGGAGTCCCC
ATCGCATGGAGGAAAAACAACACCTTCTGTCCTATCAAGTGTCCCTCAGGCAGTCAGTACAATCCATGTACCTCTGCGTGTCCTCAGC
CCAGTTGCCAGGACCCTGCAGGCTCCGGTGGCTCCTGTAATCAGCCCTGTGTGGAGGGATGTGTCTGTAATCCTGGGCTCATCCTCAG
TGGAGACAAATGTGTCCCGCTCAGTGAGTGTGGATGCACTGATGAAGGTGGAAATTACAGGCCGACAGGAGACACTTGGTTCTCAGAG
AAGGACTGTTCAGAGCGCTGTAAGTGTAACGGCAACCACAACATCACCTGCGAGCCATGGCAGTGCAGCCCTACACAGGAGTGTAAGG
TGGTGGAGGGAGTACTAGGCTGTTACTCTAGAGGAAATGGAATCTGCTCAGTATCTGGTGATCCTCACTACAACACCTTTGACAAAGT
AACCCACCACTACATGGGGTCTTGCTCCTACACCTTGACCAAACCCTGCAACGTCTCCACCGACTTGCCGTACTTCACCGTGGACACC
CAGAACGAGCACAGAGGAAGTAACAAGAGGGTTTCCTATGTCAGAGCTGTAGTGATCAATGTGAGCGGTGTGACTGTCATCCTTGGCA
AGGGACGCAAAGTCCAGGTCAATGGGACAGCAGTCGTCCCACCTTTGAATCCTGCCAAAGGAGTCAAGATCTACTTAAGTGGAAAGTT
TGTCGTCCTGGAGACAGACT

>ClS4_8_3:

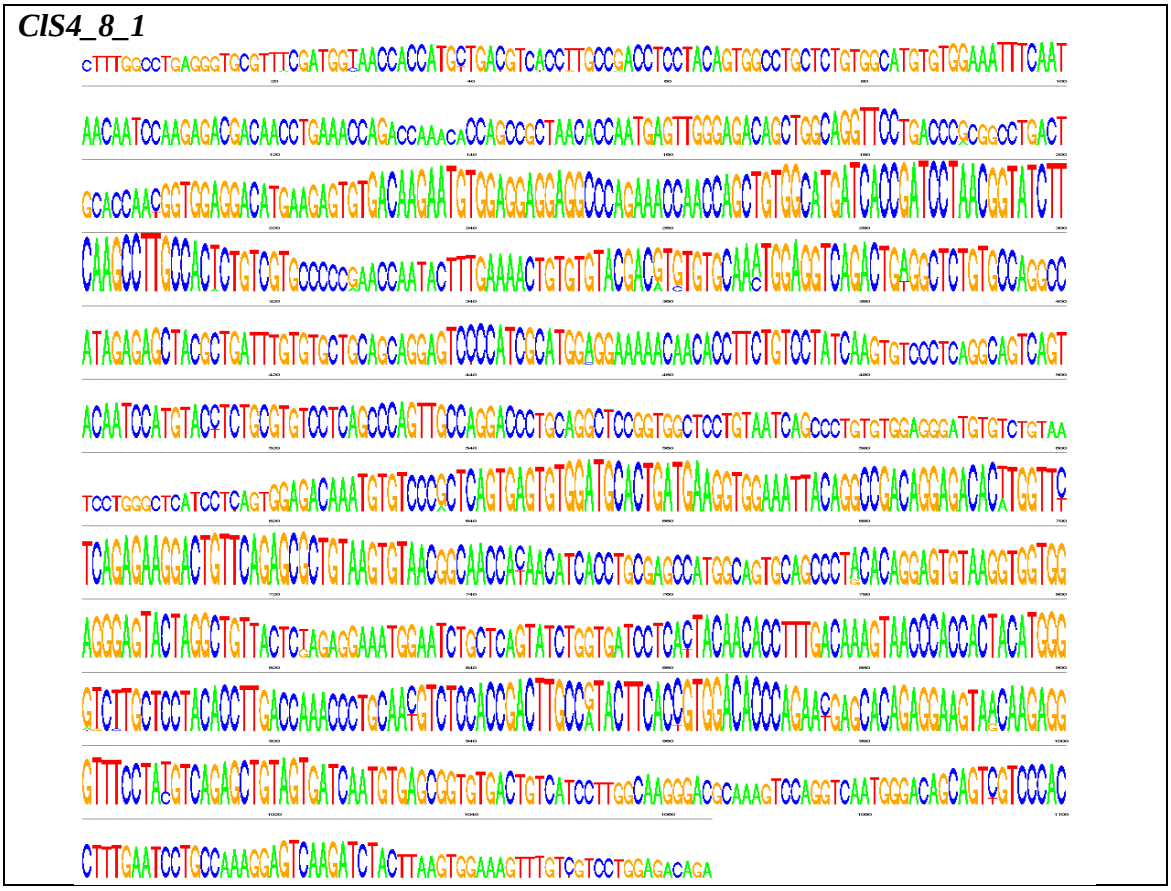
TGGCCTGAGGGTGCGTTTCGATGGTAACCACCATGCTGACGTCACCTTGCCGACCTCCTACAGTGGCCTGCTCTGTGGCATGTGTGGA
AATTTCAATAACAATCCAAGAGACGACAACCTGAAACCAGACCAAACACCAGCCGCTAACACCAATGAGTTGGGAGACAGCTGGCAGG
TTCCTGACCCGCGGCCTGACTGCACCAACGGTGGAGGACATGAAGAGTGTGACAAGAATGTGGAGGAGGAGGCCCAGAAACCAACCAG
CTGTGGCATGATCACCGATCCTAACGGTATCTTCAAGCCTTGCCACTCTGTCGTGCCCCCGAACCAATACTTTGAAAACTGTGTGTAC
GACGTGTGTGCAAATGGAGGTCAGACTGAGGCTCTGTGCCAGGCCATAGAGAGCTACGCTGATTTGTGTGCTGCAGCAGGAGTCCCCA
TCGCATGGAGGAAAAACAACACCTTCTGTCCTATCAAGTGTCCCTCAGGCAGTCAGTACAATCCATGTACCTCTGCGTGTCCTCAGCC
CAGTTGCCAGGACCCTGCAGGCTCCGGTGGCTCCTGTAATCAGCCCTGTGTGGAGGGATGTGTCTGTAATCCTGGGCTCATCCTCAGT
GGAGACAAATGTGTCCCGCTCAGTGAGTGTGGATGCACTGATGAAGGTGGAAATTACAGGCCGACAGGAGACACTTGGTTCTCAGAGA
AGGACTGTTCAGAGCGCTGTAAGTGTAACGGCAACCACAACATCACCTGCGAGCCATGGCAGTGCAGCCCTACACAGGAGTGTAAGGT
GGTGGAGGGAGTACTAGGCTGTTACTCTAGAGGAAATGGAATCTGCTCAGTATCTGGTGATCCTCACTACAACACCTTTGACAAAGTA
ACCCACCACTACATGGGGTCTTGCTCCTACACCTTGACCAAACCCTGCAACGTCTCCACCGACTTGCCGTACTTCACCGTGGACACCC
AGAACGAGCACAGAGGAAGTAACAAGAGGGTTTCCTATGTCAGAGCTGTAGTGATCAATGTGAGCGGTGTGACTGTCATCCTTGGCAA
GGGACGCAAAGTCCAGGTCAATGGGACAGCAGTCGTCCCACCTTTGAATCCTGCCAAAGGAGTCAAGATCTACTTAAGTGGAAAGTTT
GTCGTCCTGGAGACAGACTT

>ClS4_8_4:

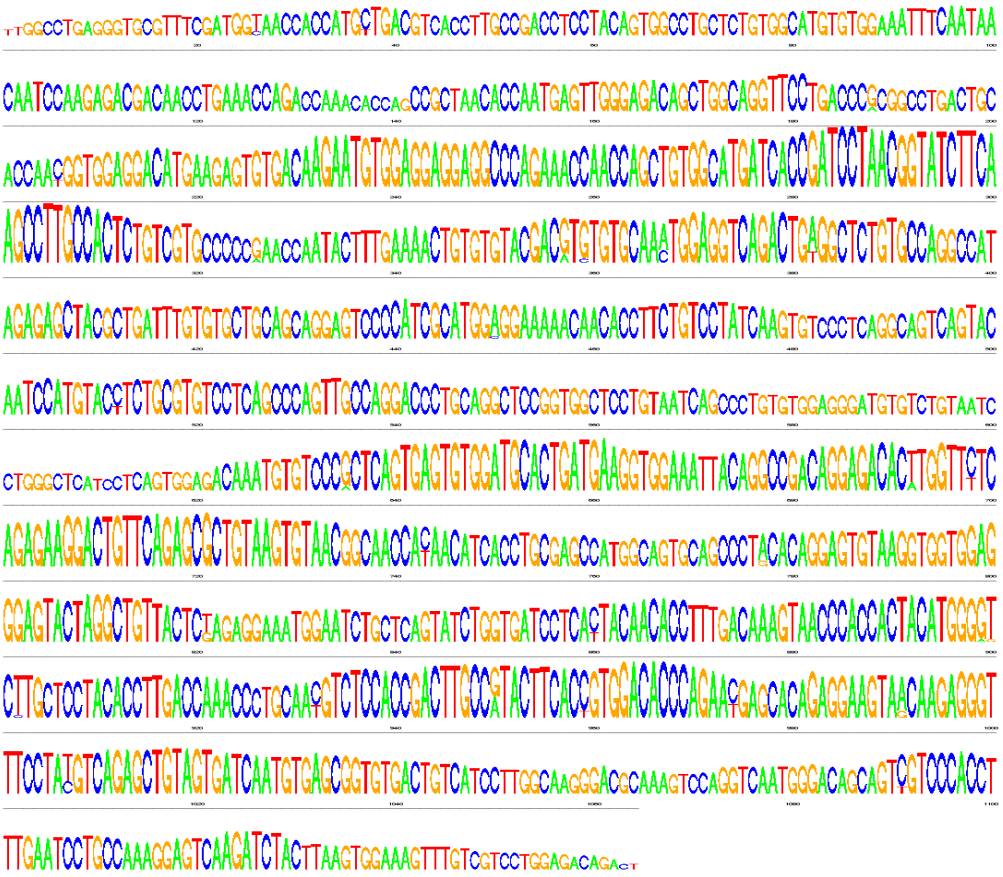
TTGGCCTGAGGGTGCGTTTCGATGGTAACCACCATGCTGACGTCACCTTGCCGACCTCCTACAGTGGCCTGCTCTGTGGCATGTGTGG
AAATTTCAATAACAATCCAAGAGACGACAACCTGAAACCAGACCAAACACCAGCCGCTAACACCAATGAGTTGGGAGACAGCTGGCAG
GTTCCTGACCCGCGGCCTGACTGCACCAACGGTGGAGGACATGAAGAGTGTGACAAGAATGTGGAGGAGGAGGCCCAGAAACCAACCA
GCTGTGGCATGATCACCGATCCTAACGGTATCTTCAAGCCTTGCCACTCTGTCGTGCCCCCGAACCAATACTTTGAAAACTGTGTGTA
CGACGTGTGTGCAAATGGAGGTCAGACTGAGGCTCTGTGCCAGGCCATAGAGAGCTACGCTGATTTGTGTGCTGCAGCAGGAGTCCCC
ATCGCATGGAGGAAAAACAACACCTTCTGTCCTATCAAGTGTCCCTCAGGCAGTCAGTACAATCCATGTACCTCTGCGTGTCCTCAGC
CCAGTTGCCAGGACCCTGCAGGCTCCGGTGGCTCCTGTAATCAGCCCTGTGTGGAGGGATGTGTCTGTAATCCTGGGCTCATCCTCAG
TGGAGACAAATGTGTCCCGCTCAGTGAGTGTGGATGCACTGATGAAGGTGGAAATTACAGGCCGACAGGAGACACTTGGTTCTCAGAG
AAGGACTGTTCAGAGCGCTGTAAGTGTAACGGCAACCACAACATCACCTGCGAGCCATGGCAGTGCAGCCCTACACAGGAGTGTAAGG
TGGTGGAGGGAGTACTAGGCTGTTACTCTAGAGGAAATGGAATCTGCTCAGTATCTGGTGATCCTCACTACAACACCTTTGACAAAGT
AACCCACCACTACATGGGGTCTTGCTCCTACACCTTGACCAAACCCTGCAACGTCTCCACCGACTTGCCGTACTTCACCGTGGACACC
CAGAACGAGCACAGAGGAAGTAACAAGAGGGTTTCCTATGTCAGAGCTGTAGTGATCAATGTGAGCGGTGTGACTGTCATCCTTGGCA
AGGGACGCAAAGTCCAGGTCAATGGGACAGCAGTCGTCCCACCTTTGAATCCTGCCAAAGGAGTCAAGATCTACTTAAGTGGAAAGTT
TGTCGTCCTGGAGACAGACT

>ClS4_8_5:

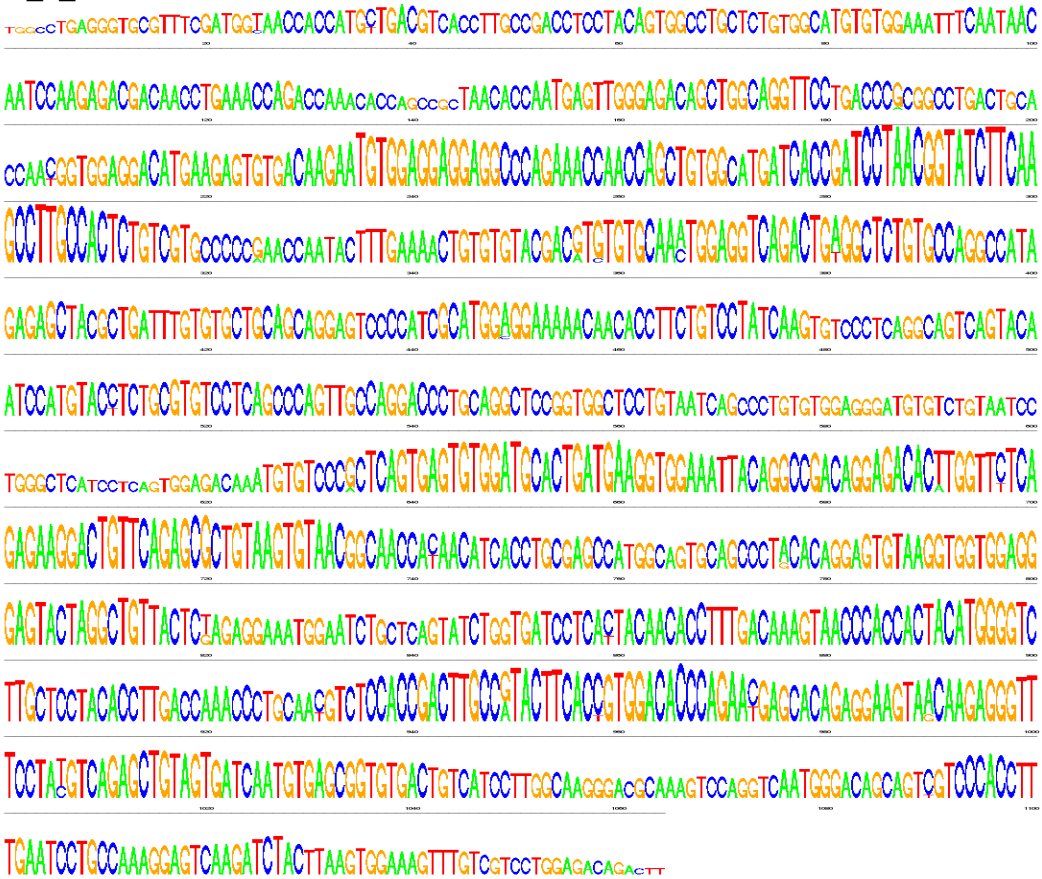
TTGGCCTGAGGGTGCGTTTCGATGGTAACCACCATGCTGACGTCACCTTGCCGACCTCCTACAGTGGCCTGCTCTGTGGCATGATCAC
CGATCCTAACGGTATCTTCAAGCCTTGCCACTCTGTCGTGCCCCCGAACCAATACTTTGAAAACTGTGTGTACGACGTGTGTGCAAAT
GGAGGTCAGACTGAGGCTCTGTGCCAGGCCATAGAGAGCTACGCTGATTTGTGTGCTGCAGCAGGAGTCCCCATCGCATGGAGGAAAA
ACAACACCTTTGACAAAGTAACCCACCACTACATGGGGTCTTGCTCCTACACCTTGACCAAACCCTGCAACGTCTCCACCGACTTGCC
GTACTTCACCGTGGACACCCAGAACGAGCACAGAGGAAGTAACAAGAGGGTTTCCTATGTCAGAGCTGTAGTGATCAATGTGAGCGGT
GTGACTGTCATCCTTGGCAAGGGACGCAAAGTCCAGGTCAATGGGACAGCAGTCGTCCCACCTTTGAATCCTGCCAAAGGAGTCAAGA
TCTACTTAAGTGGAAAGTTTGTCGTCCTGGAGACAGACT



ClS4_8_2



ClS4_8_3



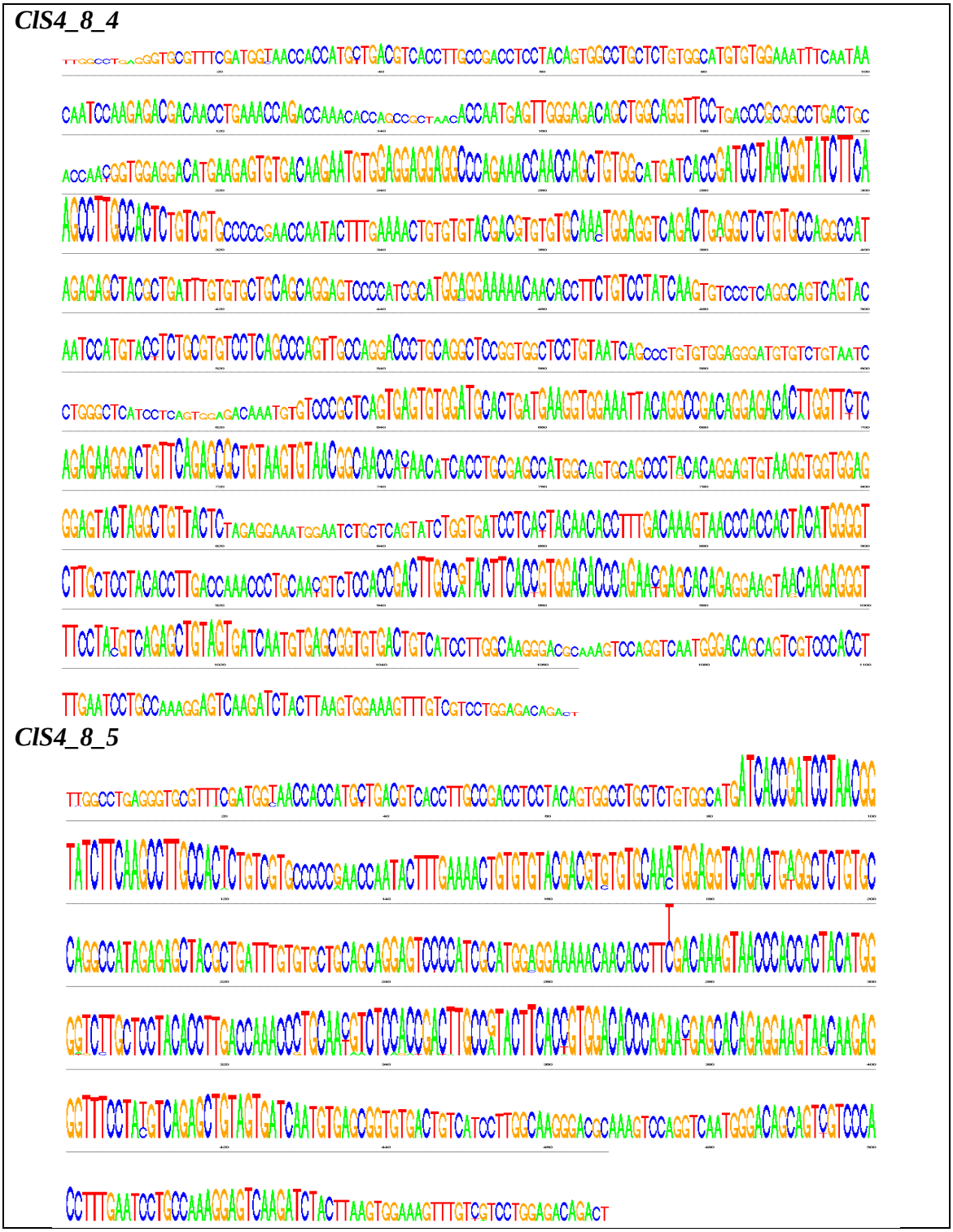


Figura III.II.VIII. Logos de los k-meros más influyentes en la construcción del cluster ClS4_8.